多模态深度学习综述

刘建伟，丁熙浩，罗雄麟

(中国石油大学(北京) 自动化系，北京 102249)

摘　要：模态是指事物发生或存在的方式，如文字、语言、声音、图形等。多模态学习是指学习多个模态中各个模态的信息，并且实现各个模态的信息的交流和转换。多模态深度学习是指建立可以完成多模态学习任务的神经网络模型。多模态学习的普遍性和深度学习的热度赋予了多模态深度学习鲜活的生命力和发展潜力。旨在多模态深度学习的发展前期，总结当前的多模态深度学习，发现在不同的多模态组合和学习目标下，多模态深度学习实现过程中的共有问题，并对共有问题进行分类，叙述解决各类问题的方法。具体来说，从涉及自然语言、视觉、听觉的多模态学习中考虑了语言翻译、事件探测、信息描述、情绪识别、声音识别和合成，以及多媒体检索等方面研究，将多模态深度学习实现过程中的共有问题分为模态表示、模态传译、模态融合和模态对齐四类，并对各问题进行子分类和论述，同时列举了为解决各问题产生的神经网络模型。最后论述了实际多模态系统，多模态深度学习研究中常用的数据集和评判标准，并展望了多模态深度学习的发展趋势。

关键词：多模态；深度学习；神经网络；模态表示；模态传译；模态融合；模态对齐

中图分类号：TP181　　doi: 10.19734/j.issn.1001-3695.2018.12.0857

Survey of multimodal deep learning

Liu Jianwei, Ding Xihao, Luo Xionglin

(*Dept. of Automation*, *China University of Petroleum*, *Beijing* 102249, *China*)

Abstract: A modality refers to the way in which something happens or is experienced, such as word, language, sound, picture and so on. Multimodality is a combination of two or more modalities. Multimodal learning refers to learning the information of each modality in the multimodality, and realizing the exchange and conversion of information of each modality. Thus, Multimodal deep learning is the establishment of a neural network model that can accomplish multimodal learning tasks. The universality of multimodal learning and the intensification of deep learning lead to the vitality of multimodal deep learning. This paper aims to summarize the current multimodal deep learning, find common problems in the implementation of multimodal deep learning under different multimodal and learning objectives, as well as making common problems classify and describing methods for solving various problems at the early development of multimodal deep learning. Specifically, this paper summarizing the current multimodal deep learning that study on natural language, visual, auditory, and considering the research direction such as language translation, event detection, information description, emotion recognition, voice recognition and synthesis, and multimedia retrieval and so on, which further concludes that there are four types of common problems: multimodal representation, multimodal interpretation, multimodal fusion, and multimodal alignment. Meanwhile, each common multimodal learning problem is sub-categorized and discussed, and the neural network models generated for solving the problems are listed. Finally, it introduce some actual multimodal system, list baseline datasets and evaluation criteria used in multimodal deep learning, and conclude with perspectives and directions for future research.

Key words: multimodal; deep learning; neural network; multimodal representation ; multimodal interpretation; multimodal fusion; multimodal alignment

0　引言

模态是指事物发生或存在的方式，多模态是指两个或者两个以上的模态的各种形式的组合。进一步解释模态和多模态的话，模态是指某种类型的信息，或者是该信息的表示；当一个研究或者数据集中包含多个模态时，它是具有多模态属性的研究或者数据集。人们听到的声音、看到的实物、闻到的味道都是一种模态，人们生活在一个多种模态相互交融的环境中。为了使人工智能更好地理解世界，必须赋予人工智能学习、理解和推理多模态信息的能力。多模态机器学习指建立模型使机器从多模态中学习各个模态的信息，并且实现各个模态的信息的交流和转换。从早期的视听语音识别研究到近期的语言和视觉模型研究，多模态机器学习在提升机器对各个模态的认知能力、加深机器对各个模态的认知深度、实现信息在机器环境下的交流互通等方面取得了显著的成效。多模态深度学习是多模态机器学习发展到现阶段的必然产物，多模态深度学习继承了之前的多模态机器学习的学习任务和学习目的，用深度学习的方法推进多模态机器学习的进步和发展，并且取得了显著的进步。

多模态机器学习研究起源于生活，也服务于生活，致力

收稿日期：2018-12-11；**修回日期：**2019-01-22

作者简介：刘建伟（1966-），男，新疆乌鲁木齐人，副研究员，博导，博士，主要研究方向为模式识别与智能系统（liujw@cup.edu.cn）；丁熙浩（1993-），男，河南安阳人，硕士研究生，主要研究方向为深度学习；罗雄麟（1963-），男，湖南汨罗人，教授，博导，博士，主要研究方向为先进过程控制．

于帮助人类解决实际问题。从 20 世纪 70 年代开始，在近几十年的研究中，多模态研究分为四个发展时期，即人类行为多模态研究、多模态计算机处理研究、多模态互动研究和多模态深度学习研究。在人类行为多模态研究时期，研究者从人类的心理行为和动作行为两方面展开研究。Mulligan 等人[1]系统地研究了人类在接受单个信号和多个信号的不同情况下，人类心理活动的表现差异；Mcgurk[2]论述了包含相同信息的模态（人类交流活动中的嘴唇动作和声音）在人类信息理解和交互中的相互作用和影响。在多模态计算机处理研究时期，研究者开始使用计算机对不同的多模态问题进行建模、训练和解决实际多模态问题。1984 年，Petajan[3]提出了第一个联合视频和声音两个模态进行语音识别的视听语音识别系统(audio-visual speech recognition, AVSR)，联合视频和声音两个模态进行声音识别的结果在原来的只有单模态声音输入的声音识别系统上实现了大幅度的飞跃，多模态机器学习开始表现出其优秀的数据学习和处理能力。AVSR 的提出和发展给多模态机器学习带来了深远的影响，自从多模态计算机处理研究时期的 1986 年实现第一个 AVSR 系统，到如今的深度学习时期，AVSR 始终在各个时期都扮演了一个先行者的角色。除此之外，在多模态计算机处理研究时期，研究者同时也致力于人机交互研究和多媒体信息的计算机处理技术研究。在人与机器交流的研究中，Krueger[4]提出了一个基于视频识别技术的人机互动沉浸式环境；Fels 等人[5]构建了一个神经网络系统，将捕捉到的手势信号转换成声音信号。在用计算机处理多媒体信息的研究中，Christel 等人[6]综合语音识别、图像理解、机器翻译等机器学习成果，使计算机能够自动地整合视频中的声音、图片和语句等各模态的信息，并生成一个包含数字视频、声音和语句的可检索的数据库。在这个阶段中，数学建模技术的发展和创新，如卷积神经网络(convolution neural network, CNN)和 BP 算法的提出、高斯混合隐马尔可夫模型的提出等，也极大地提升了计算机建模的能力，给之后的多模态深度学习建立了坚实的基础。在多模态互动研究时期，研究者通过对模态内和各模态之间的相互作用关系的学习研究，提高了各模态数据上机器学习和认知综合能力。在模拟人类的多模态学习过程方面，CALO 工程[7]以建立全新的接受多种输入信息，可以实现推理、听从命令、解释自我行为并总结经验的和帮助人处理一些计算任务的助手软件为目标，进行了多模态交互研究，并衍生出苹果手机助手 Siri 等产品；IDIAP 工程[8]开发会议助手软件，自动标注会议信息中的各模态的数据，如语音转录和会议参与者身份，帮助用户根据会议中的各模态的信息对会议档案等进行浏览，帮助用户在会议中高效获取信息。在多媒体信息检索方面，Orhan 等人[9]综合视频中的声音、图像等多模态信息，实现对视频中事件、对象交叉模态的语义检索。此外，在机器学习算法方面，研究者也进行了各种探索，如动态贝叶斯网络(dynamic Bayesian networks)和条件随机场(conditional random fields)模型等。多模态学习发展至今已进入多模态深度学习时期，近年来计算机技术和大规模数据集处理技术的迅速发展，神经网络(artificial neural networks, ANN)的高热度研究，都给深度学习带来了新的生命力和活力，刺激了深度学习在各个方面的研究和应用，多模态机器学习也在深度学习的浪潮下实现了长足的进步和发展，多模态深度学习成为了多模态机器学习的主流。在图像识别方面，自 2012 年卷积神经网络第一次在 ImageNet 数据集上展现出惊人的识别准确率后，卷积神经网络在之后的历届比赛中不断的刷新纪录，并在 2017 年最后一届 ImageNet 比赛中，取得了 top5 错误率 3.79%的超越人类的表现。神经网络对图像的强大处理能力，同时也促进着多模态深度学习在图像行为探测、图像标注、图像问答等应用场景下的表现。除计算机视觉，语音识别、机器翻译等领域也随着神经网络的引入取得了巨大突破。

多模态深度学习的发展给多模态机器学习带来了革命性的发展，使得多模态机器学习完成了巨大的飞跃。Tadas[10]总结了多模态机器学习的研究情况，提出围绕多模态机器学习，在其之后的发展中需要克服的五个挑战：模态表示、模态传译、模态对齐、模态融合和合作学习。本文旨在针对多模态深度学习的发展前期，总结当前的多模态深度学习在不同的多模态组合和学习目标下，其实现过程中的四个挑战——模态表示、模态传译、模态对齐和模态融合，合作学习（co-learning）主要围绕数据与数据之间的关系，不涉及多模态深度学习中神经网络的构造。在本文中，在自然语言、视觉、声音这三个方向上，针对各应用场景，如语音识别和生成、事件探测、图像和视频描述、面部识别和表情分析、跨媒体检索等，展开对各个挑战的论述。表 1 概括了多模态深度学习的各应用场景涉及的主要问题。表中“+”表示应用场景涉及问题类别；“+”的个数表示应用场景完成问题的难度；“-”表示应用场景未涉及问题类别。

表 1 多模态深度学习的各应用场景涉及的问题

Table 1 Challenge involved in various application scenarios of multimodal deep learning

多模态深度学习的应用	研究方向			
	模态表示	模态传译	模态对齐	模态融合
语音识别和生成				
视听语音识别	++	+	+	++
语音生成	+	++	-	-
事件探测				
图片识别	+	-	-	+
视频动作识别	++	-	-	+
情感识别				
面部表情识别	+	-	-	+
信息描述				
图像描述	+	++	++	-
视频描述	++	++	++	++
图像问答	+	+++	++	++
视觉对话	++	+++	++	+++
多媒体检索				
跨媒体检索	+	+	+	-
自然语言处理				
机器翻译	+	+++	++	+

1 模态表示

在机器学习领域，提取一个或多个模态数据的语义信息，即学习出一个或多个模态的表示，一直是一个充满挑战的问题，本文将这个问题定义为模态表示。模态表示是多模态深度学习的基础，分为单模态表示和多模态表示。单模态表示指对单个模态信息进行线性或非线性映射，产生单个模态信息的高阶语义特征表示。语句、图像、视频、声音等模态为单模态表示中的主要处理对象，且不同的模态有不同的适用的神经网络模型。多模态表示基于单模态表示，并对单模态表示的结果进行约束。多模态表示指采用模态共作用语义表示或者模态约束语义表示的方法，对各模态信息进行处理，使得包含相同或相近语义的模态信息也具有相同或相近的表

示结果。单模态表示是多模态表示的基础，且单模态表示应包含该模态输入的全部有效信息；多模态表示是单模态表示的发展，其应包含混合数据中各模态的信息。

本章从神经网络的角度对模态表示展开探讨，理解在神经网络的作用下，产生一个模态或多个模态的向量表示的过程，即模态表示在深度学习方向上的发展。

1.1 单模态表示

在本节中，为方便讨论，将单模态表示分为语句模态的表示、视觉模态的表示和声音模态的表示三种。在各个分类下进一步进行子分类，将语句模态的表示分为单词模态的独热表示、单词模态的低维空间表示、单词序列模态的袋子表示和单词序列模态的低维空间表示；将视觉模态的表示分为图像模态的表示和视频模态的表示；将声音模态的表示分为声音特征向量的提取和提取特征向量的高阶表示。

1.1.1 语句模态的表示

单词模态的独热表示。矩阵 $X=[x_1,x_2,\ldots,x_n]^T$ 表示一个句子，其中 x_i 是第 i 个单词的独热表示向量。 x_i 是一个维度等于词典包含的单词个数且元素取值为 0,1 向量，且只有一个元素值为 1，其余元素都为 0，值为 1 的元素在向量中的位置与 x_i 所表示的单词在词典中的位置坐标相同。在对语句进行单词级别的处理时，如对各个单词的词性（动词、名词等）、态度倾向（积极、消极等）和表示内容（实物、抽象概念等）等某个方面进行分类，基于单词独热表示的语句模态的表示有很好的表现。单词模态的独热表示仅把语句模态所包含的单词或字进行了简单的向量化替换，按此模态表示的进一步要求是数据语言可以反映出单词或字的语义信息。

单词模态的低维空间表示。分布性假设指一个单词或字包含的信息被其上下文中的单词确定，而不是由单词或字本身决定，例如北京、东京等首都城市名称上下文中的单词相似程度较高，这类单词或字的语义信息就相近。用 $x'=xW$ 线性方程创建一个语义空间，其中 x 为一个单词或字的独热表示向量，W 常为一个在神经网络模型上学习得到的转换矩阵，x' 是该单词或字在语义空间中的向量，在语义空间中，包含的信息相近的单词或字的表示向量距离较近[11,12]。

单词序列模态的袋子表示。单词序列指长度不定的，单词顺序明确的单词串，包括短语、句子、段落和文档。假定 x 表示一个单词序列， x 是一个维度等于词典包含的单词个数且元素取值为 0,1 的向量，值为 1 的元素在向量中的位置与单词序列中包含的所有单词在词典中的位置相同，其余元素为 0。在对单词序列进行数据处理时，袋子表示是其最基本的表示形式。袋子表示忽略了词语在单词序列中的先后顺序，考虑词语顺序后，句子袋子模型衍生出句子 n-grams 袋子模型。n-grams 袋子模型指建立 n-grams 词典，按照袋子模型的方法，产生一个维度为 n-grams 词典元素个数的表示句子的 0,1 向量。n-grams 袋子模型极大地增加了数据维度，且加剧了数据稀疏。袋子表示和 n-grams 袋子表示与单词的独热表示相似，简单且有效，并且基于袋子表示和 n-grams 袋子表示的语言模型常能获得较准确的结果，但是都没有考虑单词的语义信息。

单词序列模态的低维空间表示。单词序列模态的低维空间表示指获取单词序列模态的语义表示，即将单词序列映射到语义空间中。在早期的获取单词序列模态的语义表示的探索中，最简单的方法就是加权平均单词序列中各单词的语义表示向量，还有一种较复杂的方法是按照句子解析树的单词顺序，将句子组织为矩阵。这两种方法都有各自的缺点，前者在加权平均的过程中忽略了单词先后顺序，后者的核心是句子的解析，其适用对象局限于句子。为解决这些不足，研究者在此模型的基础上提出了多种新型且有效的模型，并在其他种类的神经网络模型上进行了尝试。基于传统的前向神经网络，Le 等人[13]提出段落向量的记忆分布模型(distributed memory model of paragraph vectors， PV-DM)，在该模型的预测过程中计算一个段落对应的段落向量。在卷积神经网络上，Kalchbrenner[14]等人用一个卷积语句模型对某个语句的 n-grams 表示进行处理，将语句的 n-grams 表示映射为由固定维度向量构成的向量序列；Zhang 等人[15]构造了一个编码器-解码器模型，在编码器部分用卷积神经网络获得了句子的向量表示，其编码器结构如图 1 所示，编码器的输入是长度为 60 的句子，X 为语句经线性映射后产生的向量序列，该向量序列被两个卷积层非线性映射为一个句子模态表示向量。在递归神经网络上，Cho 等人[16]使用单层的递归神经网络将序列单词映射为一个固定维度的隐层向量。在之后的发展过程中，Sutskever 等人[17]使用一个多层的长短记忆神经网络将一个语句映射为一个固定维度的向量；Bahdanau 等人[18]使用双向递归神经网络将一个语句编码成一个向量对序列，且每个向量对都包含了这个向量所对应单词的周围单词的信息；刘宇鹏等人[19]提出层次化递归神经网络，在底层使用训练好的循环神经网络生成包含了输入语句的短语和结构信息的词向量。综合各种模型后，由于递归神经网络的输入序列长度可变以及当前输出与之前输入有关等特性，递归神经网络成为句子模态处理中非线性映射的主流模型。

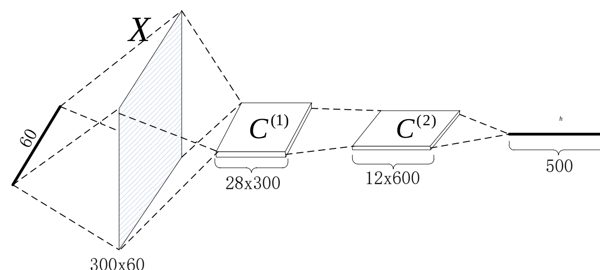


图 1　编码器结构

Fig. 1　Structure of encoder

在本节中从单词和单词序列两个方向展开论述，单词模态的独热表示和单词序列模态的袋子表示从统计的观点出发，产生了单词模态和单词序列模态的向量表示，单词模态的低维空间表示和单词序列模态的低维空间表示从探索自然语言语义的观点出发，产生了单词模态和单词序列模态的向量表示。由统计观点获得的对应模态的向量表示忽略了语句模态中固有的单词的前后顺序信息，加剧了数据稀疏，且未能提取语句的语义信息。与其相对应的语义观点则很好地解决了上述问题，对单词模态的独热表示和单词序列模态的袋子表示进行深度的语义提取，产生了低维度的、包含了对应的单词和单词序列语义信息的向量表示。

1.1.2 视觉模态的表示

视觉模态分为图像模态和视频模态，视频模态在时间维度上展开后是一个图像序列。因此，学习视觉模态的向量表示的关键问题是学习图像模态的向量表示。在图像模态表示的发展过程中，多层感知器在对图像进行处理时结果较差，不能有效地学习图像特征，并且还存在训练参数过多等问题，卷积神经网络很好地解决了这些问题，使得图像处理完成了重大突破。基于图像表示的发展，视频模态表示的研究者使用单通道卷积神经网络和双通道卷积神经网络产生了视频模态的表示。

1.1.3 图像模态的表示

在深度学习中，卷积神经网络是在多层神经网络的基础上发展起来的针对图像而特别设计的一种深度学习方法，在图像处理上取得了优异的效果。在图像特征提取的角度下，在卷积神经网络中，卷积和池化操作对图像进行特征提取，且将卷积和池化操作提取的特征矩阵输入全连接层或全局均值池化层，产生图像的特征向量。在本节中，在模态表示角度下，对经典的卷积神经网络展开论述，如 LeNet-5、AlexNet、VGG、GoogLeNet、ResNet，和 CapsNet，将卷积神经网络的卷积和池化操作理解为产生图像模态矩阵表示的过程，将全连接层或全局均值池化层的输入理解为图像模态的向量表示。

卷积神经网络 LeNet-5[20]能以极高的精度实现手写体数字和字母的识别，且应用于信封邮编识别和车牌识别中。LeNet-5 的输入为包含数字或字母的灰度图像，经过卷积和池化后产生特征图像，即图像模态的矩阵表示，特征图像经过维度变化后获得全连接层的输入，即图像模态的向量表示。LeNet-5 的高性能反映出其网络结构对包含字母和数字图片的信息的强大的提取能力，即产生包含字母或数字的图像模态的矩阵表示和向量表示的能力。

随着计算机数据处理能力的提升和深度学习技术的提升，研究者在 LeNet-5 的基础上提出了更加复杂和高效的卷积神经网络以拓展其对图像模态的特征提取能力，在近几年逐步提出了如 AlexNet、VGG、GoogLeNet、ResNet 等网络，极大地提高了图像识别的精度。与 LeNet-5 相比，AlexNet[21]通过更多的卷积和池化操作以及归一化处理和 dropout 等训练方法，在网络深层的卷积层和池化层获得图像的矩阵表示，即通过增加网络深度获取了包含图像深度语义信息的特征表示。与 AlexNet 增加神经网络深度的方式不同，VGG[22]通过构建含有多个卷积子层的卷积层实现网络深度的拓展，它每层都有 2~4 个卷积子层，用较小的卷积核和多个卷积层实现了对图片特征的精细抓取。VGG 的结构使得其能深度提取图像中的精细的语义特征，获得更好的图像模态表示。

为获得更好的图像模态表示，研究者不断地尝试增加网络的深度，但是发现当网络深度增加到一定程度后网络性能逐渐变差，获得的图像模态表示反而不能更好地提取图像模态信息。ResNet[23]是在增加网络深度的研究方向上进行了突破性探索的深度卷积网络，由融合了恒等映射和残差映射的构造性模块堆栈后构成。当在网络已经到达最优情况下继续向深层网络运算时，构造性模块中的残差映射将被置 0，只剩下恒等映射，这样使网络在更深的网络层上也处于最优。ResNet 结构简单精巧，使得随着卷积层和池化层的深度的增加，在其上的图像模态的矩阵表示所包含的语义信息不会减少，在全连接层前产生一个更加抽象的包含图像语义信息的图像模态向量表示。

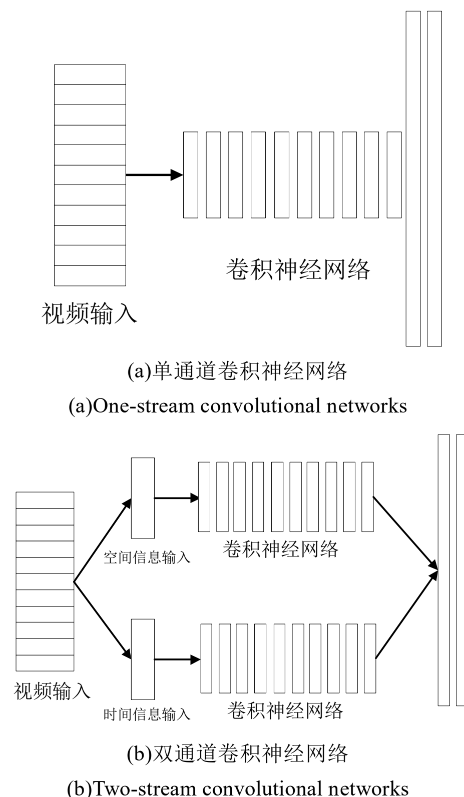
上述卷积神经网络都是使用常规的卷积和池化操作对图像进行特征提取，完成图像模态表示。除此之外，还存在一些网络对卷积或池化操作进行变形，对图像进行特征提取，完成图像模态表示。NIN(network in network)[24]提出了卷积层的改进算法 Mlpconv 层，Mlpconv 层在每个感受野中进行更加复杂的运算，获得高度非线性的图像的矩阵表示，并且 NIN 还使用全局均值池化代替全连接层，产生图像的向量表示，并提高网络的泛化能力。受 NIN 的激发，GoogLeNet[25]提出 Inception 模块。Inception 模块具有高效表达特征的能力，它包含 1x1、3x3、5x5 三种尺寸的卷积核，以及一个 3x3 的下采样，不同尺寸的卷积核赋给 Inception 模块提取不同尺寸的特征的能力。Inception 模块从纵向和横向上，增加了卷积层的深度，使得 GoogLeNet 能够产生更抽象的图像模态的矩阵表示。

尽管卷积神经网络提取的特征表示已经能够很好地包含图像中的语义信息，但是它并没有包含图像中实例的方向和空间信息，并且池化层必然会损失一些有效信息。为解决这一问题，Hinton[26]提出了 CapsNet。CapsNet 是卷积神经网络的一种拓展，它的基本组成单元是 capsule。capsule 是一组神经元，其输入和输出都是向量形式，向量中的每个元素都是图像中某个实体特征的参数表示，并且相邻的两个 capsule 层通过动态路由算法相连，实现参数选择。因此，CapsNet 能够在每个 capsule 层上产生包含图像中实例的方向和空间信息的向量表示，并且使用动态路由算法代替池化层，避免了有效信息的损失。

在本节中笔者站在模态表示的角度去理解卷积神经网络，论述了图像模态表示在卷积神经网络结构发展过程中的进步。卷积神经网络发展到当下，其提取的图像模态表示已经具有的极为抽象的语义信息。在之后的发展中，研究者可以继续增加网络的深度以提取更为抽象的语义信息，增加对卷积神经网络的结构的理解，探索其产生语义信息的过程。

1.1.4 视频模态的表示

视频为在时间维度上的图像序列，它自然地拥有空间属性和时间属性。空间属性指图像序列中每个图像包含的信息，时间属性指图像序列中相邻图像的相互作用信息。视频模态的表示应该包含视频的空间和时间两个属性信息。视频的空间属性主要由卷积神经网络提取，时间属性由卷积神经网络或长短记忆神经网络对视频中邻近的图像帧包含的运动信息提取。总结当前的视频模态的深度学习文献，根据各文献的网络结构把视频模态的表示分成单通道卷积神经网络、双通道卷积神经网络(two-stream convolutional networks, TSCN)和混合神经网络三种。图 2（a）和（b）分别表示单通道卷积神经网络和双通道卷积神经网络的结构。



(a)单通道卷积神经网络

(a)One-stream convolutional networks

(b)双通道卷积神经网络

(b)Two-stream convolutional networks

图 2　获取视频模态表示的两种神经网络结构

Fig. 2 Two neural network structures for video modal representation

单通道卷积神经网络处理对象为视频中一段连续的图像帧，它用一个卷积神经网络完成这段连续的图像中的时间和空间信息融合，并在卷积神经网络的全连接层前产生这段连续的图像帧的向量表示。单通道卷积神经网络提取视频的时空属性常有两种方式，a)改变卷积神经网络的结构，在其输入端或输出端融合视频的时间属性和空间属性；b)采用 3D 卷积核，使用卷积计算融合视频的时间属性和空间属性。在单通道卷积神经网络中，Karpathy 等人[27]将视频分片，每片都包含个数固定的在时间上邻近的多个帧，以片作为卷积神经网络的处理对象，并构造了后融合、前融合和慢融合三种用卷积神经网络在时间维度上融合片所包含的时间和空间信息的方式，使得卷积神经网络可以提取不同抽象层次的视频时空特性，在全连接层上获得视频模态的深度表示。Ji[28]和 Tran[29]采用 3D 卷积核对多个连续图像帧的堆叠形成的输入进行卷积运算，用 3D 卷积核实现对视频的时空特性的表示提取。

双通道卷积神经网络的处理对象也是视频中一段连续的图像帧，它用两个卷积神经网络分别学习这段连续的图像帧中的时间属性和空间属性，并在两个网络的全连接层前产生这段连续的图像帧的时间属性表示和空间属性表示。双通道卷积神经网络的输入包含图像帧输入和运动图像输入，图像帧输入为这段连续的图像帧中的一个图像，其包含这段图像帧中的空间信息，运动图像输入为这段连续的图像帧通过光学等技术处理产生后的向量图，如光流位移场叠加等，其包含了这段图像帧中的时间信息。

Simonyan 等人[30]首先构造了一个双通道卷积神经网络，双通道卷积神经网络由两个并行的卷积神经网络构成，其中一个神经网络的输入为视频中单个图像帧，以完成对视频的空间信息表示学习；另一个神经网络的输入为连续的多个帧的光流位移场的堆叠结果，完成对视频的时间信息的表示学习。在之后的发展中，由于图像模态表示已取得极大的突破，视频模态的研究者不断对视频的时间信息的通道添加更多的视频时间属性信息，以获取更好的时间属性表示。Wang 等人[31]基于双通道卷积神经网络，提出轨迹池化卷积映射结果，各卷积层的输出经过时空正则化或者频道正则化，输出结果再经过轨迹池化，用费舍尔向量编码轨迹池化结果，形成视频的时间信息的高维特征表示。研究者还将双通道卷积神经网络拓展到长视频的处理中，通过对长视频进行分段或抽样，产生多个连续的图像帧，将每个图像帧序列输入双通道卷积神经网络中产生该图像帧序列对应的空间和时间表示，融合每个图像帧序列的空间和时间表示后产生这段长视频的空间和时间表示。Wang 等人[32]基于双通道卷积神经网络，提出时间分段网络，对视频进行分段，将每个视频段进行时间和空间信息提取，得到各分段视频的空间和时间表示，将各空间表示和时间表示融合后获得整段视频的空间和时间表示。Wang 等人[33]对未经数据加工的视频进行动作识别研究，提出了均匀抽样或基于镜头抽样两种分片抽样策略，对视频进行抽样，使双通道卷积神经网络对每个采样视频进行处理，获得每个采样视频的空间和时间表示。

由于 LSTM 神经网络对时间序列拟合的优秀性能，研究者将 LSTM 引入到上述两种网络结构中，构造混合神经网络。混合神经网络的基本思想是将卷积神经网络的输出作为 LSTM 神经网络的输入，将视频中的图像帧或运动图像按照时间顺序依次输入卷积神经网络，卷积神经网络在每个时刻都会产生输入的图像帧或运动图像的向量表示；同时 LSTM 会读取每个时刻的图像帧或运动图像的向量表示并产生一个隐变量，该隐变量随着时间更新。在单通道卷积神经网络上，Donahue 等人[34]在卷积神经网络后添加一个双层 LSTM 网络，用卷积神经网络提取视频的图像帧中的信息，用双层 LSTM 网络学习图像帧的时间信息，产生融合了视频的空间和时间信息的隐变量。在双通道卷积神经网络上，Wu 等人[35]在双通道卷积神经网络的两个卷积神经网络的全连接层后各添加一个双层 LSTM 网络，用 LSTM 神经网络学习图像帧的空间和时间信息，以及运动图像的时间和空间信息，产生两个在不同层面融合了视频的时间和空间信息的隐变量。

本节总结了当前产生视频模态表示的神经网络的模型，按照其网络结构的不同，将其分为了单通道卷积神经网络、双通道卷积神经网络和混合神经网络，介绍并举例说明各类模型的网络结构的特点。单通道卷积神经网络和基于其的混合网络产生一个融合了视频的空间和时间信息的向量表示，单通道卷积神经网络和基于其的混合网络产生两个视频的向量表示，分别包含了其空间信息和时间信息。双通道卷积神经网络由于其对视频模态的时间和空间信息的分离式学习，其学习到的视频模态的时间和空间表示在进行视频识别等任务中具有天然的优势。因此，双通道卷积神经网络为当前获取视频模态表示的主要模型。

1.1.5 声音模态的表示

与其他信号一样，声音模态的表示就是提取声音信号的语义特征向量。在当前的包含神经网络结构的声音处理模型中，声音模态的表示主要包含两个过程：声音模拟信号转换为声音数字信号并完成特征向量的提取；提取特征向量的高阶表示。在本节根据其模型结构的不同，将提取特征向量的高阶表示的模型分为混合模型、神经网络模型、编码器-解码器模型三种结构。

1.1.6 声音特征向量的提取

声音是模拟信号，声音的时域波形只代表声压随时间变化的关系，不能很好地体现声音的特征。因此，在声音特征提取时，首先应将采集到的语音信号数字化，转换为便于计算机存储和处理的离散的数字信号序列；然后利用内含生理学、语音学相关的先验知识的数字信号处理技术对离散的数字信号序列进行声学特征向量的提取。当前的声音信号的处理技术主要有傅里叶变换、线性预测以及倒谱分析等。研究者基于这些处理技术，提取出一些当下普遍适用的声学特征，如梅尔频率倒谱系数(Mel-frequency cepstral coefficients)[36,37]、感知线性预测(perceptual linear prediction)[38]、线性预测编码(linear predictive coding)[39]和线性预测倒谱系数(linear predictive cepstral coefficients)[40]。为了进一步增强声学特征的区分性，降低模型的复杂度并提高识别效率，研究人员提出一些用于特征变换和特征降维等特征加工方法，其中代表性的方法有主分量分析(principal component analysis)[41]、线性判别分析(linear discriminant analysis)[42]和异方差线性判别分析(heteroscedastic linear discriminant analysis)[43]等。

近年来，研究者还提出了一些将特征提取和声音模型训练紧密结合在一起的方法，如利用区分性训练算法对基础特征进行变换（比较典型的有 fMPE[44, 45]、RDLT(region dependent linear transform)[46,47]）和利用不同的神经网络提取特征(比较常见的有 Tandem 特征[48, 49]、bottleneck 特征[50, 51])。

1.1.7 提取特征向量的高阶表示

在用神经网络识别声音时，提取特征向量的高阶表示是指使用神经网络对提取的声音特征向量进行多级非线性映射，学习特征向量中包含的不同抽象层次的信息。根据神经网络

在各声音识别系统中作用的不同，将其学习的不同抽象层次的信息分为以下三类。

第一种情况，在包含声音模型、语言模型和解码器的声音识别系统中，神经网络常用来与隐马尔可夫模型(hidden Markov model, HMM)组成混合结构的声音模型，称为 ANN-HMM 混合模型。其中隐含马尔可夫模型用来对声学单元和语音特征序列之间的关系建模，其隐状态为声学单元，深度神经网络对声学特征向量和隐马尔可夫模型状态的关系进行建模，即学习 HMM 状态关于给定的声音特征向量的后验概率[52]，如给定的语音特征序列中 t 时刻的特征向量 y_t，ANN 最后一层采用 softmax 函数来计算 HMM 状态 s 出现的概率：

$$p(s\mid y_t)=\frac{e^{a(s)}}{\sum_s e^{a(s)}} \tag{1}$$

其中：$a(s)$ 为状态 s 在输出层对应的输出。这种情况下，特征向量的高阶表示即为 ANN 的输出层的输出，其为特征向量的高度非线性映射的结果，且包含了该特征向量中的声学单元信息。

Bourlard 等人[53]首先将神经网络引入声音识别的声学模型中，建立了 ANN-HMM 声学模型，表现出神经网络在声音模型构建中优秀的特征表示能力。其中，神经网络的训练为有监督训练。训练数据集中每个输入帧都有状态标签，通过最小化分类错误次数的方式来训练神经网络。神经网络的引入为构造混合模型创造了更加宽广的发展空间。沿着 Bourlard 等人的思路，研究者使用各种不同的神经网络对声音进行模态表示的学习，以探索神经网络在声音模态上的表示能力。Hinton 等人[54]总结了早期的混合模型，阐述了神经网络和HMM的多种混合形式，引入深度置信网络(deep belief network, DBN)，建立了 DBN-HMM 声学模型，用深度置信网络的输出拟合给定输入关于 HMM 状态的后验概率，并论证其构造的声学模型对之后的解码过程的改进作用。Abdel-Hamid 等人[55]在常规的 ANN-HMM 模型中添加卷积层和池化层，对线性频谱和梅尔频谱进行特征提取，突出频谱中的各个频率带的特征，增加模型的鲁棒性。Sak 等人[56]构造了 RNN-HMM 模型，用 RNN 拟合声学信号的长期依赖关系。Sainath 等人[57]将 CNN、RNN 和 DNN 连接起来，构造了一个可以降低频率变化、拟合输入的依赖关系的混合网络模型。同时，隐马尔可夫模型的发展变化也改善了神经网络在混合模型中的表现，并改变了特征向量的高阶表示的内容。在声音识别的发展过程中，为拟合相邻声音信号的相互作用关系，声学单元从单音素发展到三音素，隐马尔可夫模型发展成为三音子模型。Dahl 等人[58]使用深度置信网络 DBN 学习给定的特征向量关于三音子模型中 HMM 的状态后验概率，构建了考虑相邻声音信号的相互作用关系的 CD-DNN-HMM 的声音识别模型，其结构如图 3 所示。三音子模型将声音信号的依赖关系存储到 HMM 隐状态中，这使得 DNN 学习的特征向量的高阶表示天然地包含了相邻声音信号间的依赖关系。

经过长时间的实验探索，研究者证实了混合模型对包含声音模型、语言模型和解码器的声音识别系统的促进作用，卷积神经网络能增加特征向量的高阶表示的鲁棒性，RNN 能将声音信号的依赖关系添加进特征向量的高阶表示中。但是 NN-HMM 混合模型结构复杂，且在之后的模态传译过程中，需要控制声音识别模型各部分结构对传译结果的影响程度。此外，当前常用的混合模型训练过程复杂，有多个阶段，首先训练出一个高斯混合模型—隐马尔可夫模型(Gaussian mixture model-hidden Markov model, GMM-HMM)，用该模型生成训练集中每帧声音信号的状态标签，得到显示对齐信号和标签后，组成神经网络的训练数据集；然后在生成的神经网络的训练数据集上训练神经网络，重新估计 HMM 的转移概率，并更新训练数据集。重复以上过程，直到收敛以完成神经网络和 HMM 的训练。

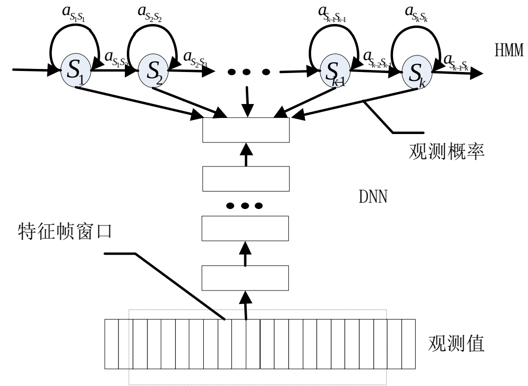


图 3 DNN-HMM 混合结构的声音模型

Fig. 3 Sound model of DNN-HMM hybrid structure

第二种情况，使用神经网络构建声音识别中的音素识别模型，即使用神经网络提取特征向量中的音素信息，获得特征向量的高度非线性映射结果，产生包含特征向量音素信息的高阶表示，实现音素识别。例如在网络的最后一层采用 softmax 函数来计算音素出现的概率，其计算结果即为给定特征向量的音素信息的高阶表示。

在早期的神经网络模型中，研究者使用设计的目标函数对神经网络进行训练，构建音素识别模型。在这个时期的神经网络的训练过程中，训练数据中每帧声音信号都有一个标注音素，即帧级对齐的训练数据。Waibel 等人[59]构造了一个时滞神经网络(time-delay neural networks, TDNN)，在网络中加入滞后单元，模拟声音信号中各帧之间的时间关系，用多层网络来学习出一个非线性的决策平面，实现音素序列识别。Graves 等人[60]使用双向的 LSTM 完成音素识别建模，其中用 LSTM 网络对声音的时间流进行建模，用双向的 LSTM 对一个帧以及其前后信息进行建模，对一个序列帧进行音素识别。

为省去数据的人为帧级对齐过程，Graves 等人[61]构造双向 LSTM 神经网络模型，定义连接主义暂态分类器(connectionist temporal classification, CTC)目标函数，采用没有帧级对齐的声音序列和音素序列对组成的训练数据，对双向 LSTM 进行训练，在训练完成后，输入要识别的信号序列，根据神经网络的输出结果，动态规划解码产生输入信号对应的音素序列。在 CTC 声音模型[62]中，声音信号的标签集合 $L'=L\cup\{blank\}$，L 由 61 个音素标签组成，blank 表示空格标签，且标签集合 L' 的元素个数为 $K=62$。CTC 声音模型中的神经网络为 RNN，神经网络的输入序列为连续的 T 帧声音特征向量序列 $\boldsymbol{x}=[x_1,\dots,x_t,\dots,x_T]$，神经网络的输出层共有 $K=62$ 个输出节点，每个输出节点都对应于标签集合中的 1 个元素 $k\in\{1,\dots,K\}$，y_t^k 为输入 x_t 产生的在输出层上第 k 个节点的输出，输入 x_t 产生的输出向量经 softmax 函数归一化处理后产生：

$$\Pr(k,t\mid \boldsymbol{x})=\frac{\exp(y_t^k)}{\sum_{k'}\exp(y_t^{k'})} \tag{2}$$

其中：$\Pr(k,t\mid \boldsymbol{x})$ 表示 t 时刻的输入向量 x_t 的分类结果，为第 k 个节点所对应的标签的概率。定义对齐向量 $\boldsymbol{a}$ 为一个长度为 T 的，$\boldsymbol{a}_t\in\{1,\dots,K\},t=1,\dots,T$ 的向量，如(1,2,15,10,62,10)，每个对齐向量都表示一个标签序列，共 K^T 个，则输入序列 $\boldsymbol{x}$ 经过

神经网络运算产生一个对齐向量 $\boldsymbol{a}$ 的概率为

$$\Pr(\boldsymbol{a}\mid\boldsymbol{x})=\prod_{t=1}^{T}\Pr(\boldsymbol{a}_t,t\mid\boldsymbol{x}) \tag{3}$$

令 B 表示将对齐向量表示的对齐结果 $(a,b,-,b,-,c)$ 和 $(a,b-,b,b,c)$ 都转录为 (a,b,b,c)。一个转录结果 $\boldsymbol{y}$ 的概率等于与 $\boldsymbol{y}$ 相对应的对齐向量的概率的加和：

$$\Pr(\boldsymbol{y}\mid\boldsymbol{x})=\sum_{\boldsymbol{a}\in B^{-1}(\boldsymbol{y})}\Pr(\boldsymbol{a}\mid\boldsymbol{x}) \tag{4}$$

令 $\boldsymbol{y}^*$ 表示输入 $\boldsymbol{x}$ 的目标转录结果，训练神经网络以最小化 CTC 为目标函数：

$$CTC(\boldsymbol{x})=-\log\Pr(\boldsymbol{y}^*\mid\boldsymbol{x})\text{。} \tag{5}$$

与混合结构的声音模型相比，神经网络构建的音素识别模型的结构和训练过程相对简单，且由于训练目标的不同，神经网络构建的音素识别模型获得的特征向量的高阶表示主要包含特征向量的音素信息，而非混合结构的声音模型获得的高阶表示包含的声学单元信息。

第三种情况，用神经网络构建编码器-解码器结构，构建声音识别模型，其中用编码器学习声音数字信号的高阶特征表示，且编码器神经网络结构中常包含 RNN 结构，使得产生的高阶特征表示中包含输入特征序列的前后帧信息[63~65]。

在本节中重点对声音信号的高阶表示模型进行论述，按模型结构不同将模型分为混合模型、神经网络模型和编码器—解码器模型三类。各模型结构的不同导致其产生的声音信号的高阶表示虽然都能很好地包含声音信号的语义信息，但其包含的语义信息各有侧重：混合结构的声音模型获得的高阶表示主要包含的声学单元信息，神经网络模型获得的高阶表示主要包含特征向量的音素信息，编码器—解码器结构则主要包含特征向量的声音的语义信息。

1.2 多模态表示

多模态表示指包含多个模态数据信息的表示，它是多个模态共用的语义空间中的向量。一个好的多模态表示应该具有平滑性、时间和空间相干性、稀疏性和自然聚类等特性。此外，Srivastava 和 Salakhutdinov 提出了多模态表示的额外理想特性：不同的多模态输入对应的多模态表示的相似性一定要反映出各多模态输入包含信息的相似性；当缺少某些模态数据信息时，依然能产生多模态表示；根据多模态表示可以获得各模态的数据信息[66]。

多模态表示基于单模态表示，并且获得多模态表示的最简单最常用的方式就是串联各模态表示。近期随着多模态研究热度的提升，获得多模态表示的方法也随之得到了迅速的发展。Tadas[10]将机器学习中多模态表示分为联合表示和协调表示，本文参考其分类结果，且根据多模态深度表示在产生多模态表示过程中各模态之间的相互作用关系和最后获得的模态表示所具有的语义信息，将多模态表示分为模态共作用语义表示和模态约束语义表示。模态共作用语义表示与联合表示的定义类似，指融合各单模态的特征表示，以获得包含各模态语义信息的多模态表示；模态约束语义表示和协调表示的定义则不相同，指用一个模态的单模态表示结果去约束其他模态的表示，以使其他模态的表示能够包含该模态的语义信息。为方便理解，用数学语言对模态共作用语义表示和模态约束语义表示进行解释。模态共作用语义表示指 $x_m=f(x_1,\dots,x_n)$，其中：x_m 为模态共作用语义表示；$x_1,\dots,x_n$ 为各模态表示；f 表示神经网络模型构建的非线性映射，模态约束语义表示指 $f(Wx_1)$，其中：W 为训练学习获得的 x_1 向量映射到 x_2 所在空间中的映射矩阵。

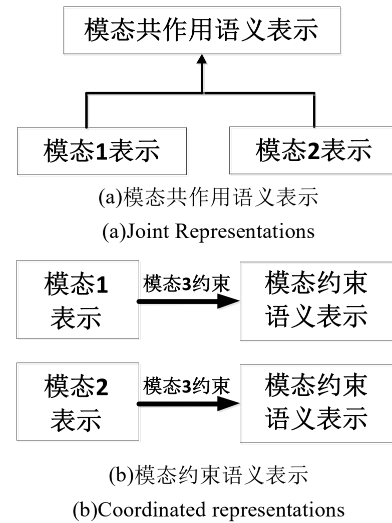


图 4 模态共作用语义表示和模态约束语义表示

Fig. 4 Joint representations and coordinated representations

1.2.1 模态共作用语义表示

深度学习中神经网络在获取自然语言、视觉、听觉等单模态表示上已经取得了卓越的成效，在单模态表示取得的发展成果上，构建更深层的神经网络以获取模态共作用语义表示。构建更深层的神经网络常用的方式为：分别用合适的神经网络学习多模态数据中的各模态数据的模态表示；然后在网络结构上继续构建深层神经网络，其输入为各模态的表示，用构建的神经网络融合各模态的语义信息获得模态共作用语义表示。由神经网络产生的模态共作用语义表示可以直接用于预测，即完成多模态的学习任务。多模态的共作用语义表示使得各模态的信息在产生多模态表示的过程中已经完成了融合，这也使得多模态的共作用语义与模态融合有了一定的交叉和相关性。为产生共作用语义表示构建的神经网络络包括前向神经网络和递归神经网络，在本节对这两种神经网络上展开论述。

在产生多模态表示的前向神经网络中，最典型的网络结构为编码器—解码器结构，其中编码器用于压缩和融合各输入模态的表示产生共作用语义表示，解码器根据产生的共作用语义表示产生学习任务的预测结果。在深层网络为编码器—解码器结构的模型中，获得各模态表示的神经网络常为经过预训练的神经网络，编码器—解码器结构的前向网络中的参数则经过端到端的训练方式产生，从而获得的多模态表示的性能优劣也可以由模型的预测结果直观地反映出。Ngiam 等人[12]构建了一个可以学习模态共作用语义表示的神经网络，并期望通过观察共作用语义表示对各模态的原始输入数据的重构能力评价共作用语义表示的性能。在整个网络模型中，首先对各输入构建解噪自编码器并完成训练，取出完成训练的解噪自编码器中的编码器作为获取各模态表示的神经网络；然后构建深层的编码器—解码器结构的前向网络，通过端到端的训练，使深层的编码器—解码器前向网络能在编码器输出层产生共作用语义表示，在解码器输出层重构各原始输入数据。根据编码器—解码器结构的特性，编码器—解码器结构中最简单的编码器可以为一个级联网络层，在该层上级联各模态的向量表示，产生共作用语义表示，这也是最基础的共作用语义表示产生方式，例如 Mroueh 等人[67]级联由神经网络学习获得的声音和视觉输入的表示，并根据级联产生的共作用语义表示产生预测结果。

仿照编码器—解码器结构，Sohn 等人[68]构建了一个深度玻尔兹曼机，以最小化各模态间的信息变化为目标，训练深

度玻尔兹曼机得到嵌入空间，且可以得到嵌入空间中各模态表示之间的联合概率分布，这使得该模型可以在某个模态缺失或损坏的情况下，根据其他模态输入得到可预测该模态信息的多模态共作用语义表示；Kim 等人[69]在视听情感识别中，采用相似的深度玻尔兹曼机融合视觉和听觉模态，产生联合表示。

递归神经网络作为上层网络产生共作用语义表示常用在预测结果受时间影响的学习任务中，如视听语音识别任务、视听情感分析。在上层递归神经网络中，递归神经单元的隐状态为融合了时间信息、各模态输入信息的共作用语义表示。在视听语音识别任务中，在每个时刻模型的底层网络会将该时刻图像和声音输入处理为图像模态表示和声音模态表示，串联两个模态表示作为上层递归神经网络的输入，此时递归神经网络的隐表示即为融合了之前各时刻图像和声音输入信息的共作用语义表示。Chung 等人[70]在此结构上作出改进，在每个底层网络的输出层添加 LSTM，使得模型底层网络获得的单模态的表示就融合了各模态的时间信息。在视听情感识别中，Chen 等人[71]在每个底层网络的输出层添加 LSTM，并依次级联各 LSTM 的输出输入到上层 LSTM 中，最上层 LSTM 的隐状态则为共作用语义表示。

在共作用语义表示模型的训练过程中，产生单模态表示的各种神经网络都可以进行预训练，或者微调经典的网络结构，整个网络的训练常采用端到端的训练方式，这也使得模型的预测结果能够反映产生的共作用语义表示是否能够充分包含各模态输入的信息。共作用语义表示训练过程简单，且能充分利用各输入包含的语义信息甚至时间信息，但是也存在训练参数过多等缺陷。

1.2.2 模态约束语义表示

模态约束语义表示不同于共作用语义表示，它不是融合各输入的信息并用于完成预测等机器学习任务，而是将输入模态的表示映射到目标模态的语义空间中，使得在目标模态表示空间中，该映射结果与语义相同的目标模态的相似性大于语义不同的目标模态，这个映射结果即为模态约束语义表示。使用神经网络获得模态约束语义表示的最主要的方法是将衡量输入模态表示和目标模态表示相似性约束条件加入目标函数中，用端到端的训练方式完成模型训练，学习获得产生输入模态和目标模态表示的神经网络的参数，以及输入模态表示映射到目标模态表示空间中的映射矩阵 W 。

模态约束语义表示思路简单，应用范围广泛，在不同的学习任务下只需要确定输入模态和目标模态的输入形式，确定适合的网络，在损失函数中添加目标模态对输入模态的约束项，就可以获得包含目标模态语义信息的输入模态约束语义表示。在图像识别问题中，Frome 等人[72]将图像表示映射到名词空间中，将铰链损失添加到损失函数中，用于对模态约束语义表示的相似性进行约束，如包含车的图像的约束语义表示和名词‘车’的损失值小于该图像与‘马’的损失值。在图像标注问题中，Kiros 等人[12]使用了相同的思想，不同的是 Kiros 用 LSTM 学习语句的表示，把图像的投影空间从名词空间拓展到了语句空间，在完成训练后使得图像在语义空间中的投影和标注语句的表示相似性最大。在跨媒体检索中，为提高检索效率，Xu 等人[73]构建了（/主语，动词，宾语/）文本语义空间而非语句空间，很好地实现了视频检索。

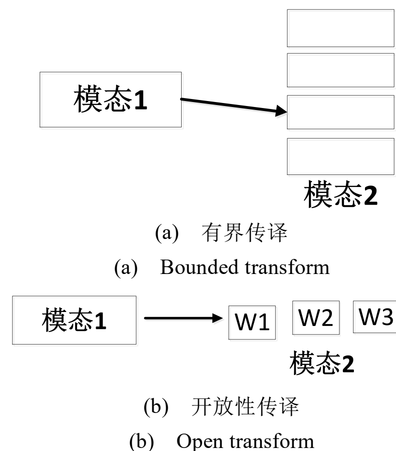
模态约束语义表示弱化了产生多模态表示过程中信息融合的必要性，采用模态约束的方式实现模态间的信息交流。模态约束语义常作为编码器出现在模型中，其输出/输入解码器产生学习任务的预测结果，在编码器完成训练后，编码器也可以对训练数据中未出现的数据类型进行编码，并投影到目标模态的语义空间中，解码器也可以对该投影结果进行处理，并产生未在训练数据中出现的预测结果。但是寻找和探索合适的包含约束的目标函数具有一定的难度，需要研究者根据各模态的特性结合实验经验构造。

模态共作用语义表示和模态约束语义表示如图 4 所示。

2 模态传译

模态传译指将模态中包含的信息传译存储在另一个模态中，实现信息在不同模态间的流通，且模态传译的研究主要集中在图片和语句、语句和声音、语言和语言等两个模态之间。模态传译是研究者长期研究的问题，很大部分的多模态深度学习都涉及模态传译，要精确地实现模态传译，模型必须能很好的理解源模态和目标模态的结构和信息。随着计算机视觉、自然语言处理和多模态数据集的发展，人们对自然语言、图像、视频、声音等模态理解程度的加深，模态传译又获得了更多的关注，且在各研究任务上取得了进一步的发展。

在本节的论述中综合考虑模态传译，按传译结果的可预测性分类，将模态传译分为有界传译和开放性传译，且分别对其展开论述。有界传译指将源模态中的一个元素传译为目标模态集合中的某个元素或多个元素。有界传译中的主要问题包括信息检索、图像识别和语音合成，如跨媒体检索、人脸识别和机器阅读等，仅需要在目标模态中找到源模态元素的对应元素，如图像识别在模态传译的概念下可以理解为：将一个包含鸟的图像传译为文字‘鸟’。开放性传译指传译结果为目标模态集合中的有前后顺序关系的多个元素组成的序列。在开放性传译研究方面，目标模态常为句子，如机器翻译、图像标注语句生成和声音识别等。在模态传译中，传译结果评价机制的主观性、源模态信号中存在的信号重复以及模态元素之间的多对一和一对多关系等问题依然为模态传译的发展中的挑战性问题。图 5 为有界传译和模态传译的示意图。图中每个方块代表一个元素。



(a) 有界传译

(a) Bounded transform

(b) 开放性传译

(b) Open transform

图 5 模态传译方式

Fig.5 Modal transform

2.1 有界传译

有界传译指将源模态中的一个元素传译为目标模态集合中的某个元素或多个元素，目标元素没有前后序列关系。跨媒体检索和语音生成为有界传译中的典型问题。

在多模态学习中，跨媒体检索指在多模态数据库中，根据给定模态类型的查询命令，检索数据库，得到另一个模态数据中包含查询命令信息的元素。多模态深度学习实现跨媒体检索的主要方式为分别学习查询模态元素和目标模态元素

的表示，然后用神经网络或者相似性评价函数学习两个模态元素的相似性，根据相似性结果完成检索，如 Feng 等人[74]构建的通信自编码器。为提升跨媒体检索的表现，研究者在更新网络结构获得更好的模态表示和更新相似性评价机制两个方面展开新的研究。在获得更好的模态表示方面，Peng 等人[75]提出一个可以同时学习媒体内和媒体间的信息，由多个神经网络堆叠形成的层次结构的跨媒体混合神经网络(cross-media multiple deep network， CMDN)，每个模态的表示由模态内和模态间信息合并生成，用前向神经网络对其相似性进行评估，实现模态检索。在更新相似性评价机制方面，Wei 等人[76]提取已完成训练的 ImageNet 的全连接层中的图像特征，使用深度语义匹配的方法将文本的的语义信息与图像特征进行配对比较，完成检索。

语音生成指构建模型学习输入模态的信息，并将其信息经声码器转换为声音输出，主要包含文字生成语音和图像生成声音。在文字生成语音的研究中，研究者通常将文本的特征向量经堆栈的双向 LSTM 处理后生成包含声音特征参数的平滑的变化轨迹，经声码器变换为声音输出。Muthukumar 等人[77]将递归神经网络当作一个训练好的传统的文本语音生成模型的后置补偿器，用其增强传统的语音生成模型的性能。Ling 等人[78]在训练阶段，用受限玻尔兹曼机和深度置信网络拟合光谱包络函数在每个隐马尔可夫状态上的分布，在声音生成阶段，根据动态特征的约束条件，遵循最大输出概率参数生成准则，使用输入语句的 RBM-HMM 或 DBN-HMM 模型，预测谱包络。在图像生成声音方面，Owens 等人[79]用卷积神经网络提取视频图像帧的图像信息，用卷积神经网络的输出作为长短记忆神经网络的输入，用长短记忆神经网络拟合视频图像帧的连续运动信息，两个神经网络共同作用，识别视频中的物体和其运动信息，再用前向神经网络合成无声视频中人们用鼓棒击打和剐蹭物体的声音。

2.2 开放性传译

开放性传译指传译结果为目标模态集合中的有前后顺序关系的多个元素组成的序列。目标模态为语句模态是开放性传译研究的主要问题，如机器翻译、图像标注、图像问答、声音识别以及手写体识别等等。

机器翻译指给出一种语言中的句子，将其翻译为另一语言中的句子。由于输入语句和输出语句的长度的可变性，以及机器翻译中单词语义的大范围依赖性，递归神经网络成为实现语句翻译的最佳神经网络模型。Cho 等人[80]用递归神经网络构造编码器-解码器，用英语短语和相应的法语短语作为平行短语对输入到递归神经网络进行训练，得到短语对的打分，将学习到的打分结果引入标准的基于短语的统计机器翻译中，改善统计机器翻译的性能。Sutskever 等人[17]使用 LSTM 将输入语句映射到高维空间中，获得输入语句的隐表示，然后使用另一个 LSTM 将输入语句的隐表示学习成为输出语句。

Luong 等人[81]在机器翻译的编码器—解码器中引入了注意力机制，构造了全局注意力模型和局部注意力模型，编码器产生一个输入语句的隐向量序列，解码器对当前要输出的目标单词隐向量与编码器输出的隐向量序列中每个向量分别求内积，求内积的数值，送入软最大函数，得到编码器输出的隐向量序列中每个隐向量的权值，编码器输出的隐向量序列经过加权形成上下文向量，上下文向量和解码器当前隐向量以及前一个时刻的解码器输出隐向量，连接组成增广向量，作为解码器的输入。当训练时，解码器的输出为使目标语言当前单词输出概率最大的网络连接边权值矩阵和偏置向量；当测试时，解码器的输出为目标语言字典中每个单词的条件后验概率，当前输出单词为字典中单词的条件后验概率最大的那个单词。在一段时间内产生输入语句的机器翻译结果。

图像识别、图像标注、图像问答和视觉对话问题为近期流行的新型研究领域，它们都是将图像模态转换为语句模态，用语句模态表示图像中所包含的信息。在各领域的研究中，研究者提出各种不同的神经网络结构以推进其发展，为适应各领域的发展，研究者也同步创建了新的多模态数据集，如 CLEVR[82]、HoME[83]、MSCOCO 等。在图像识别发展过程中，产生了多种卷积神经网络用于学习输入图像的模态表示，模态表示作为其他神经网络的输入，经过学习，输出相应的识别结果，LeNet、CapsNet 将数字手写体图像传译为图片中的数字，AlexNet、GoogLeNet 高精度地识别出 ImageNet 图片集中图片包含的物体信息。RCNN 首先对输入图像进行处理，生成 1~2 千个候选区域，将每个候选区域输入卷积神经网络对其进行特征提取，提取的特征送入每一类的 SVM 分类器，判断是否属于该类，实现了图片多目标检测和识别。图像语句标注和图像问答在图像识别技术基础上迅速发展。Vinyals 等人[84]采用了编码器-解码器结构，用 GoogLeNet 作为编码器，生成图像的固定长度向量表示，使用 LSTM 作为解码器，将向量解码为语句，整个编码器—解码器将图片转换为描述图片内容的语句。图像问答系统能够根据图像回答与图像相关的问题，图像问答旨在评估图片标注训练结果的好坏。Antol 等人[85]使用卷积神经网络学习图像特征，使用 LSTM 学习提问语句信息，将两个学习结果输入一个前向神经网络，取经软最大处理后的几个最优的输出作为回答。

在图像问答中，由于语句天然会包含其叙述对象的相关信息，提问语句作为输入会给模型提供问题语句中的先验信息，进而会导致模型不能真实地理解图片中包含的信息。为解决这一问题，Johnson 等人[82]创建了用于诊断图像问答模型能否在理解图片的基础上给出提问语句相应回答的数据集 CLEVR；Hu 等人[86]提出了端到端的模块网络（end-to-end module networks， N2NMNs），它能够直接从文本输入预测新的模块化网络体系结构，并将其应用到图像中，以解决问答任务，并且这个模型在 CLEVR 数据集上取得了很好的效果，表现出它在图像问答中充分考虑图像信息的能力；Santoro 等人[87]提出了关系网络（relation networks， RN），它能够有效地实现关系推理，即充分利用提问语句和图像的关系，作出回答。

视觉对话是近期新兴的多模态任务，它由图像标注和图像问答综合发展而来，其任务为实现人与机器使用自然语言交流视觉数据，将视觉模态信息传译到语句模态中。Abhishek 等人[88]首先提出一个模型结构为编码器—解码器神经网络的视觉对话系统，编码器部分使用后融合编码器、分层递归编码器和记忆网络编码器对图像信息和提问者的询问信息进行编码，产生图像信息和历史问答信息的公用语义表示，解码器部分使用生成式解码器和判别式解码器产生回答语句。之后，Abhishek 等人[89]通过构建一个混合结构的神经网络催生一个虚拟的智能体，该智能体能充分理解 3D 图像包含的各种信息，并基于 3D 图像，结合问题产生回答语句。

声音识别将语音模态转换为语句模态。声音识别的传统过程包括建立声学模型、语言模型和解码过程。解码指根据声学模型和语言模型，将输入的语音特征向量序列转换为字符序列，实现模态传译。在当前的研究中，基于 HMM 模型的语音识别系统一般基于加权有限状态转换器（weighted finite state transducer）进行解码，将语音识别的解码问题归结为加权有限状态转换器的最优路径搜索问题。在搜索最优

路径时，常用启发式的柱搜索技术[90]。此外，一些研究者也提出通过最优化解码图中的路径成本总和来最小化最终解码误差。Mohri 等人[91]对 CTC 目标函数进行改进，在完成神经网络的训练后，输入要识别的信号，根据神经网络的输出结果，解码产生与输入信号对应的单词序列。在构造编码器—解码器模型的声音识别中，编码器—解码器模型共同学习一个语音识别的所有过程，实现模态传译。

在本节将模态传译分为有界传译和开放性传译，并对两种模态传译方式分别展开讨论。完成模态传译的过程常和解决学习任务的过程同步进行，这也反映出模态传译和机器学习任务之间的关系，即模态传译为涉及模态间信息交流的机器学习任务的抽象概括。在有界传译和开放性传译的讨论中，本文列举各典型的学习任务，并分析该学习任务中常用的神经网络的结构，以帮助理解神经网络在模态传译中的功能，展示深度学习在模态传译方面的发展。

3　模态融合

多模态融合指综合来自两个或多个模态的信息以进行预测的过程。在预测的过程中，单个模态通常不能包含产生精确的预测结果所需的全部有效信息，多模态融合过程融合了来自两个或多个模态的信息，实现信息补充，拓宽输入数据所包含信息的覆盖范围，提升预测结果的精度，提高预测模型的鲁棒性[92]。多模态融合按多模态融合与各模态建模的先后关系分为前融合、后融合和混合融合。前融合指在模态建模之前，通过集成或组合来自所有模态的特征来完成特征层面的融合[93]；后融合指分别执行每种模态的建模，然后综合模型的输出或决策以产生最终决策结果，完成决策层面的融合[94]；混合融合指在特征级别和决策级别进行融合，组合前融合和后融合的方法[95]。图 6 为各融合方式的结构示意图。在多模态深度学习中，神经网络的结构可以直观地反映该网络实现模态融合的方式。最早使用神经网络实现模态融合的多模态任务为视听语音识别，现在它的范围已经拓展到了图像问答、视觉对话、手势识别、情感分析以及视频识别和描述。

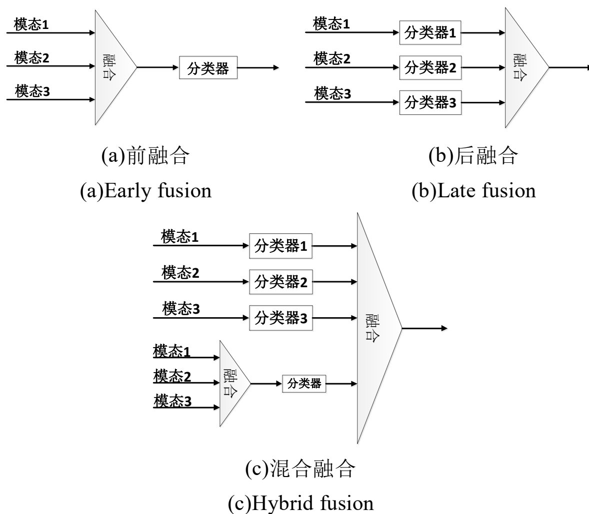


图 6　模态融合方式

Fig. 6　Modal fusion

3.1　前融合

前融合实现过程中，首先提取各输入模态的特征；然后将提取的特征合并到被称为特征集成的特征集合中，集成的特征作为输入数据输入到一个模型中，输出预测结果。前融合具有较低的计算复杂性，但是各模态特征经转换和缩放处理后产生的集合特征通常具有较高的维度，可以使用主成分分析(PCA)和线性判别分析(LDA)对集合特征进行降维处理[96]。前融合产生的特征集成和模态共作用语义表示，两者的区别模糊不清，并且模态共作用语义表示也可以作为前融合特征集成结果。与其他两种融合方式相比，前融合因其简单的结构和训练过程，常出现在各多模态学习任务中，并且使用前融合表示产生的预测结果常作为该预测任务的基准。

前融合中模态表示的融合有多种方式，常用的方式有对各模态表示进行相同位置元素的相乘或相加、构建编码器—解码器结构和用 LSTM 神经网络进行信息整合。在输入模态为脑电图和眼睛信号的情绪识别任务中，Liu 等人[97]使用双峰自编码神经网络对提取的脑电图和眼睛信号进行特征集成和重构；然后将集成特征输入支持向量机，产生情绪分类识别结果。在图像问答任务中，Antol 等人[93]对卷积神经网络和 LSTM 神经网络学习到的图像和问题的向量表示中对应元素相乘。在视频描述任务中，Donahue 等人[34]在卷积神经网络上叠加 LSTM 神经网络，实现了对视频中的时间和空间信息的融合。

3.2　后融合

后融合过程先提取各模态的特征，将提取的各模态特征输入对应的模型中，每个模型输出一个预测结果；然后整合各模型的预测结果，形成最终的预测结果。与前融合相比，后融合可较简单地处理数据的异步性，系统可以随模态个数的增加进行扩展，每个模态的预测模型能更好地对该模态进行建模，当模型输入缺少某些模态时也可以进行预测。同时，后融合也存在一些缺点，如未考虑特征层面的模态相关性、后融合实现难度更高等。

在后融合过程中，整合各模型的预测结果的常用方式为平均[98]、投票[99]、基于信道噪声和信号方差的加权[100]和模型选择（如 Adaboost[101]和神经网络）。在视频识别或视频描述任务中，由于视频模态经过双通道卷积神经网络后，产生视频的时间信息表示和空间信息表示，所以在产生识别结果和描述语句前，神经网络需融合视频的时间和空间信息。在视频识别任务中，Simonyan 等人[30]分别将视频的时间信息表示和空间信息表示输入全连接层，且全连接层的最后一层为 softmax 函数，之后采用平均和支持向量机两种方法对 softmax 结果进行融合，获得视频预测结果；在视频描述任务中，Pan 等人[102]使用 LSTM 神经网络融合由卷积神经网络获取的与视频相关的单词信息，生成视频描述语句。

3.3　混合融合

混合融合是组合了前融合和后融合的方法，其在综合了前融合和后融合的优点的同时，也增加了预测模型的结构复杂度和训练难度。

由于神经网络结构的多样性和灵活性，在实现混合融合的研究中，神经网络得到了广泛的应用。Wu 等人[95]构建了视频和声音信号经过仅基于视频信号和仅基于声音信号的听声辨人模型，产生模型预测，同时视频信号和声音信号的集成特征输入视听相关模型(audio-visual correlative model，AVCM)，产生模型预测，采用加权方式整合模型预测，获得识别结果。在图像问答任务中，Xu[103]和 Lu[104]用递归神经网络和卷积神经网络分别学习问题语句和图像的信息，用注意力机制实现问题语句信息和图像信息的融合。在手势识别任务中，Neverova 等人[105]分别用卷积神经网络对手势视频中的左手、右手包含的时间信息和空间信息进行学习和融合，并行地使用卷积神经网络提取身体姿势的图像信息和声音信息；然后使用全连接神经网络进行各信息融合，在输出层输

出辨识结果。

在本节中按多模态融合与各模态建模的先后关系将模态融合分为前融合、后融合和混合融合，分别讨论各种融合方式的优缺点，并用现有模型解释各融合方式模型的构建方式。同时，研究者经过在不同的研究任务下对比各融合方式后，发现各融合方式并无确定的优劣关系，在不同的实验条件下，研究者可以尝试不同的融合方式以获得更好的模态融合结果。

4　模态对齐

多模态对齐指辨别来自两个或两个以上的不同模态的元素之间的关系。例如，在机器翻译中，寻找存在于‘I am a Chinese’和‘我是中国人’这两个不同语言模态的句子中的‘I-我’，‘am-是’，‘a Chinese-中国人’的对齐关系；在图像标注中，给出一个图像和对应的标注语句，辨别标注语句中与图像各区域对应的单词或短语。在多模态深度学习中，本文根据对齐算法实现对齐的方式，将模态对齐分为注意力对齐和语义对齐。注意力对齐综合考虑输入模态中各元素与目标模态中某个元素的关系，实现模态对齐；语义对齐根据输入模态各元素与目标模态中各元素语义相似性，构建语义对齐数据集，在语义对齐数据集上训练模型，最终使得学习得模型能够自然地实现语义对齐。这两种方式都在模态对齐中取得了较好的成果，其中注意力对齐模态元素更能较好地考虑模态元素之间的长期依赖关系。

4.1　注意力对齐

在一个包含注意力对齐的深度学习模型中，注意力对齐的主要功能为学习输出中某个元素与输入中各元素的对齐概率。注意力对齐应用范围广泛，如机器翻译[18,81]、图像标注[106]、语音识别[64,65]等涉及模态传译的多模态学习任务，并在各个领域取得了良好的表现，且在深度学习模型中加入注意力机制常能提升该模型的性能。

注意力对齐作用过程复杂，本文以机器翻译中的软注意力模型为例介绍注意力模型的构造过程。Bahdanau 等人[18]采用解码器—编码器结构实现注意力对齐和机器翻译，其模型结构示意图如图 7 所示。在编码器阶段，使用 BRNN 输入语句中的单词进行语义特征提取，令 T_x 为输入文本长度，$h=\{h_j\}, j\in 1,\dots,T_x$ 为各输入时刻 j 的 BRNN 隐变量的集合。在解码器端，训练时，使用 RNN 和多层神经网络求解在每个时刻 i 使输出单词 y_i 的条件后验概率最大的模型参数；测试时，使用 RNN 和多层神经网络，把当前时刻 i 输出的所有单词中，条件后验概率最大的单词 y_i 作为当前时刻的翻译结果。令 s_i 表示解码器输出单词 y_i 在 RNN 中的隐变量，引入注意力机制，建立注意力对齐模型，该对齐模型由两层神经元构成的神经网络构建，输出表示为

$$e_{ij}=a(s_{i-1},h_j)=v_a^T\tanh(W_a s_{i-1}+U_a h_j) \tag{6}$$

其中：v_a，W_a 和 U_a 为权值矩阵。注意力对齐模型在生成一个预测单词时，能够计算输入语句中每个单词的表示与输出语句中该单词对应的预测目标单词的相关性强弱的能量值 e_{ij}，并将 e_{ij} 作为软最大函数输入，计算得到权值 $a_{ij}=\frac{\exp(e_{ij})}{\sum_{k=1}^{T_x}\exp(e_{ik})}$。使用对齐模型产生的权值和编码器产生的隐变量集合，计算解码器各输出时刻的单词的隐变量 s_i 的上下文向量 $c_i=\sum_{j=1}^{T_x}a_{ij}h_j$，并将 i 时刻上下文向量 c_i 和解码器 RNN 中的隐变量 s_i 以及 $i-1$ 时刻的输出单词 y_{i-1} 输入解码器 RNN，产生 i 时刻输出单词 y_i 的条件后验概率 $p(y_i|\{y_1,\dots,y_{i-1}\},c_i)=g(y_{i-1},s_i,c_i)$。在这个模型中，输入语句中的每个单词都以对应的概率对 i 时刻的输出单词进行对齐，将这个模型称为软注意力模型。

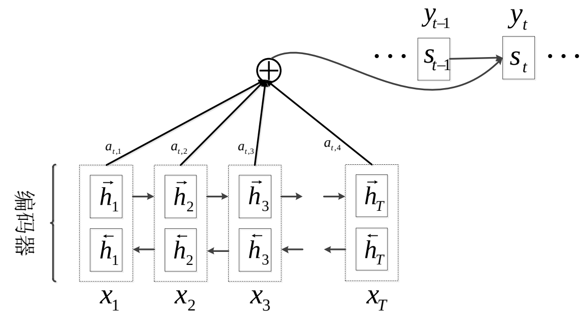


图 7　软注意力模型

Fig. 7　Soft attention model

注意力模型在发展过程中，其模型不断更新，以软注意力模型为基础发展出了硬注意力模型、全局注意力模型与局部注意力模型，以及静态注意力模型和动态注意力模型。

与软注意力模型相对应的是硬注意力模型[107]，其常用于图片文字标注中，首先使用编码器将输入图像转换为多个向量，每个向量对应于图像的一个区域，选取某个向量与目标句子单词对齐，其他的向量与目标句子单词对齐的概率硬性地置 0。

使用软注意力模型和硬注意力模型的对齐思想，考虑输入句子中注意力分配概率所覆盖单词范围大小，并且对软注意力模型进行简化和推广后，Luong 等人[81]提出全局注意力模型与局部注意力模型，并用于机器翻译。全局注意力模型与局部注意力模型的分类依据为在生成上下文向量时，是否使用源语句中所有的单词的隐表示。

在使用软注意力的机器翻译模型中，对目标句子中的每一个单词，都需要计算与这个单词相对应的源输入句子中每个单词的对齐概率，因此也称软注意力模型为动态注意力模型。Hermann 等人[108]相对应地提出了静态注意力模型，对于整个目标句子，整体对源输入句子求出一个注意力概率分布上下文向量，然后用于问答系统;梁斌等人[109]用学习获得的词向量注意力矩阵、词性注意力矩阵和位置注意力矩阵与输入词向量进行线性运算，得到经注意力产生的语句矩阵。

4.2　语义对齐

语义对齐是一种直接赋给模型对齐能力的对齐方式，语义对齐最主要的实现方式就是处理带有标签的数据集并产生语义对齐数据集，用深度学习模型去学习语义对齐数据集中的语义对齐信息[51]。

在视觉模态和语句模态对齐方面，由于带有对齐标签的视觉和语句模态数据集大小的爆发式增加，基于深度学习的有监督语义对齐算法取得了很大的进步。在图像标注中，Karpathy 等人[110]提出了如图 8 所示的神经网络模型，用包含语义信息的目标函数对数据集进行训练，然后根据训练好的神经网络和新构建的链式结构的马尔可夫随机场，动态的最小化能量函数寻找最好的语义对齐的图像和语句或单词对，并构建语义对齐的数据集。在这个过程中，Dahl 使用 RCNN 对图像进行区域划分，选取最优的 19 个区域和整个图像，共 20 个图像，使用 CNN 对 20 个图像分别进行处理，产生图像特征表示；使用 BRNN 对描述语句中的单词进行语义特征提取，其维度与图像表示维度相同；假设有 k 个图像，l 个描述语句，v_i 表示图像第 i 个局部区域表示向量，s_t 为描述语句中的第 t 个单词向量，乘积 $v_i^T s_t$ 表示两者的对齐分数，计算每个图像区域和每个单词的对齐分数，取每个单词关于各图

像区域的最大对齐分数和每个图像区域关于各单词的最大对齐分数，令图片对语句的对齐分数 S_{kl} 为取得的每个单词的最大对齐分数的加和，语句对图片的匹配分数 S_{lk} 为取得的每个图像的最大对齐分数的加和，构造目标函数

$$C(\theta)=\sum_k\left[\sum_l \max(0,S_{kl}-S_{kk}+1)+\sum_l \max(0,S_{lk}-S_{kk}+1)\right] \tag{7}$$

目标函数使神经网络输出的最优对齐的图像—语句对，比其他对齐的图像—语句对有更高的对齐分数，完成 CNN 和 BRNN 神经网络的训练；然后，沿着图像标注句子，构造一个链式结构的马尔可夫随机场，最小化能量函数后，输出对齐的区域图像和语句片段，并存储在数据集中。

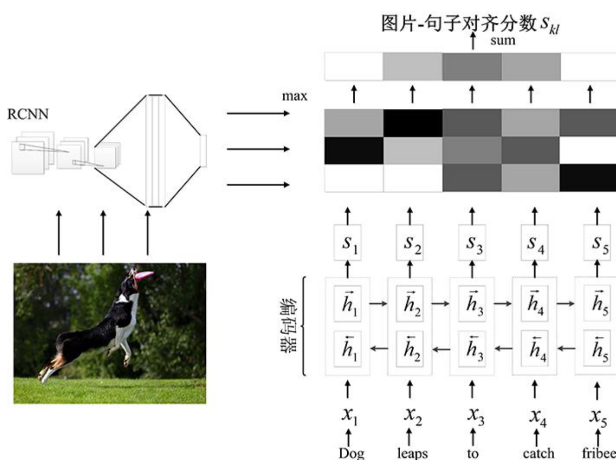


图 8　语义对齐

Fig. 8　Semantic alignment

除此之外，还有很多研究者在视觉模态和语句模态对齐方面进行了其他方式的尝试。Zhu 等人[111]通过训练一个卷积神经网络来评价电影场景和剧本段落的相似性；Mao 等人[112]使用一个卷积神经网络视觉模型和一个 LSTM 神经网络语言模型评估图像中的实例和其指称表达之间的匹配度；Yu 等人[113]在 Mao 的工作基础上，在模型中添加图像实例的外形信息和指称表达包含的上下文信息，减少错误评估。

在本节中将模态对齐分为注意力对齐和语义对齐，并总结了当前实现注意力对齐和语义对齐的方法。注意力对齐动态地使用概率对齐实现模态对齐，使得模型能够从众多输入信息中按概率比率提取信息，进而输出预测结果。语义对齐则通过探索带标签的数据集中，标签与数据之间的子元素对齐信息，构建静态的语义对齐数据集，并通过构建模型学习语义对齐信息，获得能够产生包含语义对齐信息输出的模型。两种对齐方式相比，在结构上，注意力对齐模型结构简单，形式灵活；在训练过程中，注意力对齐模型中超参数和模型参数相对较少，训练难度低；在预测结果上，注意力对齐能更好的考虑到模态元素之间的长期依赖关系，但是语义对齐能够产生语义对齐数据集，有着直观的评测结果。在实际使用中，注意力对齐由于其优势和较好的性能表现，它更频繁地出现在了各学习任务中。

5　实际多模态系统

多模态深度学习应用范围广泛，在语音识别和生成、图像识别、事件监测、情感分析和跨媒体检索等方面均有应用，它可以赋予机器理解和融合图像、语言、文字、视频等模态所包含信息的能力，具有巨大的商业价值。这吸引了很多商业公司对多模态深度学习进行商品开发，使得多模态深度学习走进实际生活。在本节中对谷歌、微软、苹果、Facebook 等科技公司的多模态应用进行论述。

谷歌公司是全球最大的搜索引擎公司，同时它也在引领世界人工智能的发展，其旗下的 DeepMind 更是人工智能领域的明星。WaveNet 是一种新型的深度神经网络，它能够根据文本产生比现有技术更好、更逼真的语音，其改进版本在 Google 智能助理中为美式英语、日语和印度尼西亚语生成逼真的声音。谷歌翻译可提供 80 种语言之间的即时翻译，支持两种语言之间的子词、语句和网页翻译，谷歌翻译手机 App 更是支持相机拍摄翻译和 11 种手写语言的翻译，实现了图像和文字、语言和语言模态信息的交流互通。Google Lens 是一款基于图像识别和光学字符识别技术的人工智能应用，它能够让机器学会“看图说话”，实现图像模态和文字模态信息的转换，也可以完成信息检索，例如可以识别图像中的实例，输出实例名称；扫描公司或商店的外观，调出公司或商店的详细信息和评价。Google Duplex 实现人工智能与人类使用自然语言自然流畅的交流，如它可以自己给饭馆、理发店等商店打电话帮助用户预订时间，并且可以产生“嗯哼”类情感助词的回答，使得人工智能更加的“像人”。Google Photos 利用人工智能技术分析相片内容自动的给相片添加标签，让用户可以使用内置的编辑工具轻松修复照片，也可以利用人工智能自动创建拼贴、动画、电影、风格化图片等。Google Allo 是一款人工智能短信应用，它能够根据用户历史输入，了解用户的对话习惯，根据接受的图像或文字短信，自动给出回复建议。Google Assistant 是一款融合了 Google Lens、WaveNet、谷歌翻译等谷歌公司前沿机器学习技术的十分强大的人工智能助手，它有着先沿的自然语言处理能力，可以与用户实现对话、文字交流等信息交互，并理解用户指令调用其他软件或硬件，也可以理解用户输入的图像或视频，识别并分析该视频内容，帮助用户认知视频中的各种信息。

Facebook 是一家世界级的社交网络服务公司，引领着机器学习在社交软件上的应用和发展。Facebook 中机器学习的作用主要是让网络社交变得更加有趣、方便，提升用户的体验度。视频风格渲染和图像风格渲染将视频或图片的风格艺术化为其他艺术风格，例如将一个真实图片渲染为梵高作品艺术风格的图片，使得视频和图片更加活泼多样；文字翻译能将朋友圈文字内容和评论内容翻译为用户设置的语言；自动给视频添加隐藏式字幕，帮助用户理解视频信息；为盲人生成图像说明，识别盲人用户浏览的图像，生成文字说明并朗读文字；人脸识别能自动识别图像中出现的人，并标注任务名称；自动检测并删除不良内容，减少社交网络中的不良信息，构建健康和谐的网络环境；内容推送可以根据用户的浏览习惯自动推送用户感兴趣的图像或视频。

特斯拉为一家电动车及能源公司，其创造的自动驾驶技术是一个典型的多模态应用系统。自动驾驶技术是通过电脑系统实现无人驾驶汽车的技术，其输入主要有视觉信息、雷达信息、全球定位系统的位置信息、语音信息、自然语言信息等。无人驾驶技术会自动识别视觉信息中的车道标志线、行人和汽车，语音信息和自然语言信息中的驾驶人命令，雷达信息中的车辆、行人、障碍物，然后综合各信息确定当前汽车的行驶状态，并决定汽车之后的行驶方向和速度。

苹果、微软、亚马逊、腾讯、百度、阿里巴巴等公司也应用了各种机器学习算法构建了大同小异的多模态应用系统，如苹果、微软、亚马逊的人工智能助手 Siri、Cortana、Alexa；腾讯的图像识别和标注；阿里巴巴淘宝的商品推荐系统；百

度的自动驾驶；可以理解语言并与人类互动、可以识别人类、使用面部自然表情甚至与人进行眼神交流取得公民资格的机器人索菲亚。

除此之外，多模态系统还应用于导航、生理病变研究、环境监测、天气预报、安全监控等领域，如生物医学图像识别中的 CT(computed tomography)技术；用图像识别技术对航空遥感和卫星遥感图像通常用图像识别技术进行加工，提取有用信息，进行天气预报和环境监测等；采用图像识别技术实现人脸识别、指纹识别、车牌识别，提高社会安全水平。

6 国内外多模态深度学习公用数据集

多模态深度学习作为一个有着极大发展潜力的深度学习的研究方向，大量的研究机构在对其现有的模型不断地进行创新和探索，完善数据集，提高多模态深度学习模型运算速度，提高输出预测准确率。在本章列举常见的多模态任务和其相应的数据集，并列出学习任务的学习情况。

表 2 汇总了各多模态深度学习问题和其相应的数据集，及基于该问题和相应的数据库学习结果。

表 2 多模态深度学习问题分类、常用的数据集和评价标准

Table 2 Multimodal deep learning problem, data sets and evaluation criteria

多模态深度学习问题分类	常用数据集	数据集介绍	最优学习结果
人脸识别[114]	LFW	包含 13K 张图片，每个图片平均包含 2.3 个人脸	ACC：99.78%
	MegaFace	包含 4.7M 张图片，每个图片平均包含 100 个人脸	ACC：86.47%
	SLLFW	包含 13K 张图片，每个图片平均包含 2.3 个人脸	ACC：85.78%
	CACD	包含 163K 张图片，每个图片平均包含 87.7 个人脸	ACC：98.50%
	IJB-A	包含 25K 张图片，每个图片平均包含 11.4 个人脸	ACC：98.20%
面部表情识别[115]	CK+	包含 123 个人的 593 个脸部图像序列	ACC：98.60%
	MMI	包含 75 个人的脸部高分辨率图像和 2900 个脸部图像序列	ACC：78.53%
	FER2013	包含 35887 张人脸图像	ACC：75.10%
	SFEW 2.0	包含 1635 张人脸图像	ACC：61.60%
	TED	包含 4178 张人脸图像	ACC：88.90%
图像标注[116]	Flcikr8k	包含 8000 张图片，每个图片有 5 个注释语句	BLUE-1:0.670；BLUE-2:0.459；
	Flcikr30k	包含 31783 张图片，每个图片有 5 个注释语句	BLUE-1:0.669；BLUE-2:0.462；
	MSCOCO	包含 123287 张图片，每个图片有 5 个注释语句	BLUE-1:0.724；BLUE-2:0.555；
图像问答[117]	DAQUAR	包含 1449 张室内场景图片，每个图片都有问答语句对	ACC：46.13%；WUPS@0.9：51.83
	COCO-QA	包含 123287 张图片，每个图片都有一个问答语句对	ACC：70.98%；WUPS@0.9：78.35
	HoME	包含 45000 张 3D 图像，每个图像都有描述语句和复杂的标注	ACC：35.8%
	CLEVR	包含 100000 张图像，每个图像都有问答语句对	ACC：55%
视听语音识别[118]	M2VTS	包含 25 个男性和 12 个女性讲话的声音信号的视频信号	ACC：96.57%
	TULIPS1	包含 7 个男性和 5 个女性讲话的声音信号的视频信号	EER：1.74
	VidTIMIT	包含 24 个男性和 19 个女性讲话的声音信号的视频信号	EER：5.23
	CUAVE	包含 19 个男性和 17 个女性讲话的声音信号的视频信号	ACC：95%
	XM2VTS	包含未知性别分布的共 295 人讲话的声音信号的视频信号	ACC：89%
机器翻译[119]	WMT'14	包含多种欧洲语言的两两语言的语句对	BLEU：41.62
跨媒体检索[120]	Wikipedia	包含 2866 个文本图像对的文档集，每个文本图像对都标有相应的语义类别	MAP：0.3608
	NUS-WIDE	包括来自 Flickr 的 269648 张图片和相关标签，总共有 5018 个独特标签	MAP：0.365
	NUS-WIDE-10k	为 NUS-WIDE 的一个子集	MAP：0.374
	Pascal Sentences	包括 1000 个图像，每个图像对应有 5 个标注语句	MAP：0.334

*ACC：准确率（Accuracy， ACC）；BLUE-n：标注语句评价指标；WUPS：吴-帕尔默相似(Wu-Palmer similarity， WUPS)，EER：等错误率（equal error rate， EER），MAP：平均精度均值（Mean average precision， MAP）

7 多模态深度学习的发展方向

a)提出关于神经网络的完备的数学描述和理论体系。神经网络为实现多模态深度学习的主要工具，神经网络的理论体系的成熟，定能给多模态深度学习带来更多的实现手段和进步。

b)构建大型多模态数据库，充分发挥深度学习技术在多模态数据集上的学习能力。深度学习的学习效果常取决于数据库所包含的信息，好的数据库可以使神经网络充分学习各种知识，避免神经网络的过拟合等问题。

c)探索更精细的模态数据特征表示，不断减小语义相同的不同模态的数据在语义嵌入空间中距离。多模态表示中，解决语义鸿沟，实现各模态信息的无障碍的交流互通，为其主要目标，探索更好的语义嵌入空间，使得多模态数据在语义空间上实现更好的信息交流。

d)参数量化分析，探索简洁的参数形式和高效的训练算法。多模态深度学习的模型参数个数往往非常多，以至于限制了多模态深度学习的应用场景。拓展神经网络的结构形式，发现高效的训练算法，实验比较和理论分析神经网络处理各种模态的能力也是摆在研究者面前的主要挑战。

e)赋予机器学习数据库外的模态能力，即模态泛化能力，在已有模态上学习的多模态表示和多模态模型能够推广到未见模态上。再完善的数据库也不能拥有全部知识，让机器具有高效、准确的学习数据库外数据的能力，是多模态深度学

习的必然产物。

f)多模态学习中的各种神经网络结构的组合形式，具有人为选择任意性，没有一个统一的标准，以便判定这种组合形式的好坏。多模态学习的模态表示学习也没有一个统一的标准，到底是怎样把模态组合起来，是一个从理论到具体算法实践亟待解决的问题。

g)多模态深度学习的目标函数通常为非凸优化问题，目前的深度学习训练算法不能避免鞍点问题，导致寻优过程失败，使得研究者无法知道到底是优化过程没有找到最优解使得预测结果不好，还是其他的模态表示和模态组合有问题。应该尽快提出求解非凸优化问题的优化求解算法。

8 结束语

在深度学习飞速发展的当下，人工智能逐渐走上历史舞台，而赋予机器接受、综合、处理各种外界信息，并对接受的信息作出反映，则是对人工智能的基本要求。多模态深度学习则为实现该基本要求的一种有效的手段。本文总结了多模态深度学习的现状，对深度学习在多模态学习中模态表示、模态传译、模态融合以及模态对齐方面的应用进行了总结。模态表示是多模态深度学习的基础，指将模态所包含的信息以什么数据形式存储在电脑中。模态表示分为单模态表示和多模态表示，语句、视觉和声音等模态为单模态表示的主要处理对象，多模态表示基于单模态表示，按照其融合模态信息的方式分为了模态共作用语义表示和模态约束语义表示。模态传译指将模态中包含的信息传译存储在另一个模态中，按照传译结果的可预测性分类。模态传译分为有界传译和开放性传译。有界传译指将源模态中的一个元素传译为目标模态集合中的某个元素或多个元素，开放性传译指传译结果为目标模态集合中的有前后顺序关系的多个元素组成的序列。模态融合指综合来自两个或多个模态的信息以进行预测的过程。模态融合按照信息融合的方式，分为前融合、后融合和混合融合。模态对齐指辨别来自两个或两个以上的不同模态的元素之间的关系。模态对齐中，常用的两种方式为注意力对齐模态元素和语义对齐模态元素。多模态深度学习中，模态表示、模态传译、模态融合和模态对齐这四个方面的研究进度并不相同。模态融合已经经过了较长时间的研究，但是近期在模态表示、模态传译和模态对齐上的研究也促进了大量的新的多模态算法的产生，并且拓展了多模态学习的应用范围。作为一种能让机器拥有更多人类智能特性的学习方法，多模态深度学习定能在之后的一个时期获得长足的发展。

参考文献：

[1] Mulligan R M, Shaw M L. Multimodal signal detection: independent decisions vs. integration [J]. Percept Psychophys, 1980, 28 (5): 471-478.

[2] Mcgurk H, Macdonald J. Hearing lips and seeing voices [J]. Nature, 1976, 264 (5588): 746-748.

[3] Petajan E D. Automatic lipreading to enhance speech recognition (speech reading) [D]. Illinois, USA: University of Illinois at Urbana-Champaign, 1984.

[4] Krueger M W. Artificial reality II [M]. New Jersey, USA: Addison-Wesley, 1991.

[5] Fels S S, Hinton G E. Glove-talk: a neural network interface between a data-glove and a speech synthesizer [J]. IEEE Trans on Neural Netw, 1993, 4 (1): 2-8.

[6] Christel M, Stevens S, Wactlar H. Informedia digital video library [J]. Communications of the ACM, 1995, 38 (4): 57-58.

[7] Tur G, Stolcke A, Voss L, *et al.* The CALO meeting assistant system [J]. IEEE Trans on Audio Speech & Language Processing, 2010, 18 (6): 1601-1611.

[8] Moore D. The IDIAP smart meeting room. martigny [R/OL]. http://glat.info/ma/av16. 3/com02-07. pdf.

[9] Orhan O, Hochreiter J, Poock J, *et al.* University of central florida at TRECVID 2008 content based copy detection and surveillance event detection [C]/ Proc of the Trecvid 2008 Workshop Participants Notebook Papers. Berlin, Germany: Springer Press, 2008.

[10] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: a survey and taxonomy [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2017, 41 (2): 423-443.

[11] Li Hang, Lu Zhengdong. Deep learning for information retrieval [C]// Proc of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. New York: ACM Press, 2016: 1203-1206.

[12] Kiros R, Salakhutdinov R, Zemel R S. Unifying visual-semantic embeddings with multimodal neural language models [EB/OL]. (2014-11-10) [2018-6-24]. https://arxiv. org/abs/1411. 2539.

[13] Le Q V, Mikolov T. Distributed representations of sentences and documents [C]// Proc of the 31th International Conference on Machine Learning. Cambridge, MA: MIT Press, 2014: 1188-1196.

[14] Kalchbrenner N, Blunsom P. Recurrent continuous translation models [C]// Proc of the 2013 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL Press, 2013: 1700-1709.

[15] Zhang Yizhe, Shen Dinghan, Wang Guoyin, *et al.* Deconvolutional paragraph representation learning [C]// Proc of Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 4172-4182.

[16] Cho K, Merrienboer B V, Gulcehre C, *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]// Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL Press, 2014: 1724-1734.

[17] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C]// Proc of Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 3104-3112.

[18] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [EB/OL]. (2016-05-19) [2018-06-24]. https://arxiv. org/abs/1409. 0473.

[19] 刘宇鹏，马春光，张亚楠. 深度递归的层次化机器翻译模型 [J]. 计算机学报，2017, 40 (4): 861-871. (Liu Yupeng, Ma Chunguang, Zhang Yanan. Hierarchical machine translation model based on deep recursive neural network [J]. Chinese Journal of Computers, 2017, 40 (4): 861-871.)

[20] Lecun Y. LeNet-5, convolutional neural networks [EB/OL]. (2010-03-03) [2018-6-24]. http://yann. lecun. com/exdb/lenet.

[21] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of the 26th Annual Conference on Neural Information Processing Systems 2012. Cambridge, MA: MIT Press, 2012: 1106-1114.

[22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10) [2018-06-24]. https://arxiv. org/abs/1409. 1556.

[23] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Deep residual learning for image recognition [C]// Proc of IEEE Conference on

Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 770-778.

[24] Lin Min, Chen Qiang, Yan Shuicheng. Network in network [EB/OL]. (2014-03-04) [2018-06-24]. https://arxiv. org/abs/1312. 4400.

[25] Szegedy C, Liu Wei, Jia Yangqing, *et al.* Going deeper with convolutions [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 1-9.

[26] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules [C]// Proc of Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 3859-3869.

[27] Karpathy A, Toderici G, Shetty S, *et al.* Large-scale video classification with convolutional neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 1725-1732.

[28] Ji Shuiwang, Xu Wei, Yang Ming, *et al.* 3D convolutional neural networks for human action recognition [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2012, 35 (1): 221-231.

[29] Tran D, Bourdev L, Fergus R, *et al.* C3D: Generic features for video analysis [EB/OL]. (2015-10-07) [2018-06-24]. https://arxiv. org/abs/1412. 0767.

[30] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [C]// Proc of Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 568-576.

[31] Wang Limin, Qiao Yu, Tang Xiaoou. Action recognition with trajectory-pooled deep-convolutional descriptors [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 4305-4314.

[32] Wang Limin, Xiong Yuanjun, Wang Zhe, *et al.* Temporal segment networks: towards good practices for deep action recognition [J]. ACM Trans on Information Systems, 2016, 22 (1): 20-36.

[33] Wang Limin, Xiong Yuanjun, Lin Dahua, *et al.* UntrimmedNets for weakly supervised action recognition and detection [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 6402-6411.

[34] Donahue J, Hendricks L A, Rohrbach M, *et al.* Long-term recurrent convolutional networks for visual recognition and description [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 2625-2634.

[35] Wu Zuxuan, Wang Xi, Jiang Yugang, *et al.* Modeling spatial-temporal clues in a hybrid deep learning framework for video classification [C]// Proc of the 23rd Annual ACM Conference on Multimedia Conference. New York, NY: ACM Press, 2015: 461-470.

[36] Zheng Fang, Zhang Guoliang, Song Zhanjiang. Comparison of different implementation of MFCC [J]. Journal of Computer Science and Technology, 2001, 16 (6): 582-589.

[37] Matusov E, Kanthak S, Ney H. On the integration of speech recognition and statistical machine translation [C]// Proc of the 9th European Conference on Speech Communication and Technology. 2005: 3177-3180.

[38] Hermansky H. Perceptual linear predictive (PLP) analysis of speech [J]. Journal of the Acoustical Society of America, 1990, 87 (4): 1738-1752.

[39] Mccree A V, Barnwell T P I. A mixed excitation LPC vocoder model for low bit rate speech coding [J]. IEEE Trans on Speech & Audio Processing, 1995, 3 (4): 242-250.

[40] Haeb-Umbach R, Ney H. Linear discriminant analysis for improved large vocabulary continuous speech recognition [C]// Proc of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE Press, 1992, 1: 13-16.

[41] Viszlay P, Pleva M, Juhár J. Dimension reduction with principal component analysis applied to speech supervectors [J]. Journal of Electrical & Electronics Engineering, 2011, 4: 245-250.

[42] Martin A, Charlet D, Mauuary L. Robust speech/non-speech detection using LDA applied to MFCC [C]// Proc of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE Press, 2001, 1: 237-240.

[43] Kumar N, Andreou A G. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition [J]. Speech Communication, 1998, 26 (4): 283-297.

[44] Povey D. Improvements to fMPE for discriminative training of features [C]// Proc of Eurospeech, European Conference on Speech Communication and Technology. Lisbon, Portugal: ISCA, 2005: 2977-2980.

[45] Povey D, Kingsbury B, Mangu L, *et al.* fMPE: discriminatively trained features for speech recognition [C]// Proc of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE Press, 2005: 961-964.

[46] Zhang Bing, Matsoukas S, Schwartz R. Discriminatively trained region dependent feature transforms for speech recognition [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2006: 313-316.

[47] Xu Jian, Yan Zhijie, Huo Qiang. A comparative study of fMPE and RDLT approaches to LVCSR [C]// Proc of International Symposium on Chinese Spoken Language Processing. Piscataway, NJ: IEEE Press, 2013: 21-24.

[48] Hermansky H, Ellis D P W, Sharma S. Tandem connectionist feature extraction for conventional HMM systems [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2000, 3: 1635-1638.

[49] Sharma S, Ellis D, Kajarekar S, *et al.* Feature extraction using non-linear transformation for robust speech recognition on the Aurora database [C]// Proc of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE Press, 2000: 1117-1120.

[50] Yu Dong, Seltzer M. Improved bottleneck features using pretrained deep neural networks [C]// Proc of International Speech Communication Association. Lisbon, Portugal: ISCA, 2011: 237-240.

[51] Macherey W. Discriminative training and acoustic modeling for automatic speech recognition [C]// Proc of the 8th European Conference on Speech Communication and Technology. Geneva, Switzerland: ISCA, 2003.

[52] 侯一民，周慧琼，王政一. 深度学习在语音识别中的研究进展综述 [J]. 计算机应用研究，2017，34 (8): 2241-2246. (Hou Yiming, Zhou Huiqiong, Wang Zhengyi. Overview of speech recognition based on deeplearning [J]. Application Research of Computers, 2017, 34 (8): 2241-2246.)

[53] Bourlard H A, Morgan N. Connectionist speech recognition [C]// Proc of the 4th European Conference on Speech Communication and Technology. Madrid, Spain: ISCA, 1995..

[54] Hinton G, Deng Li, Yu Dong, *et al.* Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups [J]. IEEE Signal Processing Magazine, 2012, 29 (6): 82-97.

[55] Abdel-Hamid O, Mohamed A R, Jiang H, *et al.* Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2012: 4277-4280.

[56] Sak H, Vinyals O, Heigold G, *et al.* Sequence discriminative distributed training of long short-term memory recurrent neural networks [C]// Proc of the 15th Annual Conference of the International Speech Communication Association. Singapore, Singapore: ISCA, 2014: 1209-1213.

[57] Sainath T N, Vinyals O, Senior A, *et al.* Convolutional, long short-term memory, fully connected deep neural networks [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2015: 4580-4584.

[58] Dahl G E, Yu Dong, Deng Li, *et al.* Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition [J]. IEEE Trans on Audio Speech & Language Processing, 2012, 20 (1): 30-42.

[59] Ieee A W M, Hanazawa T, Hinton G, *et al.* Phoneme recognition using time-delay neural networks[J]. Readings in Speech Recognition, 1990, 1 (2): 393-404.

[60] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Netw, 2005, 18 (5-6): 602-610.

[61] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2013: 6645-6649.

[62] Graves A. Supervised sequence labelling with recurrent neural networks [M]. New York: Springer, 2012: 1-131.

[63] Chorowski J, Bahdanau D, Cho K, *et al.* End-to-end continuous speech recognition using attention-based recurrent nn: first results [EB/OL]. (2014-12-04) [2018-06-24]. https://arxiv. org/abs/1412. 1602.

[64] Chan W, Jaitly N, Le Q V, *et al.* Listen, Attend and spell: a neural network for large vocabulary conversational speech recognition [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2016: 4960-4964.

[65] Chorowski J, Bahdanau D, Serdyuk D, *et al.* Attention-based models for speech recognition [J]. Future Generation Computer Systems, 2015, 10 (4): 429-439.

[66] Srivastava N, Salakhutdinov R. Multimodal learning with deep Boltzmann machines [J]. Journal of Machine Learning Research, 2014, 15 (1): 2949-2980.

[67] Mroueh Y, Marcheret E, Goel V. Deep multi-modal learning for audio-visual speech recognition [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2015: 2130-2134.

[68] Sohn K, Shang Wenling, Lee H. Improved multimodal deep learning with variation of information [C]// Proc of Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 2141-2149.

[69] Kim Y, Lee H, Provost E M. Deep learning for robust feature generation in audiovisual emotion recognition [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2013: 3687-3691.

[70] Chung J S, Senior A, Vinyals O, *et al.* Lip reading sentences in the wild [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 3444-3453.

[71] Chen Shizhe, Jin Qin. Multi-modal dimensional emotion recognition using recurrent neural networks [C]// Proc of International Workshop on Audio//visual Emotion Challenge. New York, NY: ACM Press, 2015: 49-56

[72] Frome A, Corrado G S, Shlens J, *et al.* DeViSE: a deep visual-semantic embedding model [C]// Proc of the 27th Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 2121-2129.

[73] Xu Ran, Xiong Caiming, Chen Wei, *et al.* Jointly modeling deep video and compositional text to bridge vision and language in a unified framework [C]// Proc of the 29th AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2015: 2346-2352.

[74] Feng Fangxiang, Wang Xiaojie, Li Ruifan. Cross-modal retrieval with correspondence autoencoder [C]// Proc of ACM International Conference on Multimedia. New York, NY: ACM Press, 2014: 7-16.

[75] Peng Yuxin, Huang Xin, Qi Jinwei. Cross-media shared representation by hierarchical learning with multiple deep networks [C]// Proc of the 25th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Press, 2016: 3846-3853.

[76] Wei Yunchao, Zhao Yao, Lu Canyi, *et al.* Cross-modal retrieval with cnn visual features: a new baseline [J]. IEEE Trans on Cybernetics, 2017, 47 (2): 449-460.

[77] Muthukumar P K, Black A W. Recurrent neural network postfilters for statistical parametric speech synthesis [EB/OL]. (2016-01-26) . https://arxiv. org/abs/1601. 07215.

[78] Ling Zhenhua, Deng Li, Yu Dong. Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis [J]. IEEE Trans on Audio Speech & Language Processing, 2013, 21 (10): 2129-2139.

[79] Owens A, Isola P, Mcdermott J, *et al.* Visually indicated sounds [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 2405-2413.

[80] Cho K, Merrienboer B V, Gulcehre C, *et al.* Learning phrase representations using rnn encoder-decoder for statistical machine translation [C]// Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL Press, 2014: 1724-1734.

[81] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation [C]// Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL Press, 2015: 1412-1421

[82] Johnson J, Hariharan B, Maaten L V D, *et al.* CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 1988-1997.

[83] Brodeur S, Perez E, Anand A, *et al.* HoME: a household multimodal environment [EB/OL]. (2017-11-29) [2018-06-24]. https://arxiv. org/abs/1711. 11017.

[84] Vinyals O, Toshev A, Bengio S, *et al.* Show and tell: a neural image caption generator [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 3156-3164.

[85] Antol S, Agrawal A, Lu Jiasen, *et al.* VQA: visual question answering [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 2425-2433.

[86] Hu Ronghang, Andreas J, Rohrbach M, *et al.* Learning to reason: end-to-end module networks for visual question answering [C]// Proc of

IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 804-813.

[87] Santoro A, Raposo D, Barrett D G T, *et al.* A simple neural network module for relational reasoning [C]// Proc of Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 4974-4983.

[88] Das A, Kottur S, Gupta K, *et al.* Visual Dialog [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 1080-1089.

[89] Das A, Datta S, Gkioxari G, *et al.* Embodied question answering [EB/OL]. (2017-12-01) . https://arxiv. org/abs/1711. 11543.

[90] Graves A. Sequence transduction with recurrent neural networks [EB/OL]. (2012-11-14) [2018-06-24]. https://arxiv. org/abs/1211. 3711.

[91] Mohri M, Pereira F, Riley M. Weighted finite-state transducers in speech recognition [J]. Proc Isca Automatic Speech Recognition, 2002, 16 (1): 69-88.

[92] Lahat D, Adali T, Jutten C. Multimodal data fusion: an overview of methods, challenges, and prospects [J]. Proceedings of the IEEE, 2015, 103 (9): 1449-1477.

[93] Nefian A V, Liang Luhong, Pi Xiaobo, *et al.* Dynamic Bayesian networks for audio-visual speech recognition [J]. Eurasip Journal on Advances in Signal Processing, 2002, 2002 (11): 1-15.

[94] Snoek C G M, Worring M, Smeulders A W M. Early versus late fusion in semantic video analysis [C]// Proc of the 13th ACM International Conference on Multimedia. New York, NY: ACM Press, 2005: 399-402.

[95] Wu Zhiyong, Cai Lianhong, Meng H M. Multi-level fusion of audio and visual features for speaker identification [C]// Proc of International Conference on Biometrics. Piscataway, NJ: IEEE Press, 2006: 493-499.

[96] Atrey P K, Hossain M A, Saddik A E, *et al.* Multimodal fusion for multimedia analysis: a survey [J]. Multimedia Systems, 2010, 16 (6): 345-379.

[97] Liu Wei, Zheng Weilong, Lu Baoliang. Emotion recognition using multimodal deep learning [C]// Proc of the 23rd International Conference on Neural Information Processing. New York: Springer, 2016: 521-529.

[98] Shutova E, Kiela D, Maillard J. Black holes and white rabbits: metaphor identification with visual features [C]// Proc of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL Press, 2016: 160-170.

[99] Morvant E, Habrard A, Ayache S. Majority vote of diverse classifiers for late fusion [C]// Proc of Structural, Syntactic, and Statistical Pattern Recognition. New York: Springer, 2014: 153-162.

[100] Evangelopoulos G, Zlatintsi A, Potamianos A, *et al.* Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention [J]. IEEE Trans on Multimedia, 2013, 15 (7): 1553-1568.

[101] Liu Qingju, Wang Wenwu, Jackson P. A visual voice activity detection method with adaboosting [C]// Proc of the Sensor Signal Processing for Defence. Piscataway, NJ: IEEE Press, 2012: 1-5.

[102] Pan Yingwei, Yao Ting, Li Houqiang, *et al.* Video captioning with transferred semantic attributes [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 984-992.

[103] Xu Huijuan, Saenko K. Ask, attend and answer: exploring question-guided spatial attention for visual question answering [C]// Proc of the 14th European Conference on Computer Vision. Berlin, Germany: Springer press, 2015: 451-466.

[104] Pan Lu, Lei Ji, Wei Zhang, *et al.* R-VQA: learning visual relation facts with semantic attention for visual question answering [EB/OL]. (2018-1-20) [2018-06-24]. https://arxiv. org/abs/1805. 09701.

[105] Neverova N, Wolf C, Taylor G, *et al.* ModDrop: adaptive multi-modal gesture recognition [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2016, 38 (8): 1692-1706.

[106] Xu K, Ba J, Kiros R, *et al.* Show, attend and tell: neural image caption generation with visual attention [C]// Proc of the 32nd International Conference on Machine Learning. New York: ACM Press, 2015: 2048-2057.

[107] Meutzner H, Ma Ning, Nickel R, *et al.* Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2017: 5320-5324.

[108] Hermann K M, Kočiský T, Grefenstette E, *et al.* Teaching machines to read and comprehend [C]// Proc of Annual Conference on Neural Information. Cambridge, MA: MIT Press, 2015: 1693-1701.

[109] 梁斌，刘全，徐进，等. 基于多注意力卷积神经网络的特定目标情感分析 [J]. 计算机研究与发展，2017，54 (8)：1724-1735. (Liang Bing, Liu Quan, Xu jin, *et al.* Aspec-t based sentiment analysis based on multi-attention CNN [J]. Journal of Computer Research and Development, 2017, 54 (8): 1724-1735.)

[110] Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 3128-3137.

[111] Zhu Yukun, Kiros R, Zemel R, *et al.* Aligning books and movies: towards story-like visual explanations by watching movies and reading books [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2016: 19-27.

[112] Mao Junhua, Huang J, Toshev A, *et al.* Generation and comprehension of unambiguous object descriptions [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015, 2 (1): 11-20.

[113] Yu Licheng, Poirson P, Yang S, han *et al.* Modeling context in referring expressions [C]// Proc of the 14th European Conference on Computer Vision. Berlin, Germany: Springer Press, 2016: 69-85.

[114] Wang Mei, Deng Weihong. Deep face recognition: a survey [EB/OL]. (2018-9-28) [2018-06-24]. https://arxiv. org/abs/1804. 06655.

[115] Li Shan, Deng Weihong. Deep facial expression recognition: a survey [EB/OL]. (2018-10-22) [2018-06-24]. https://arxiv. org/abs/1804. 08348.

[116] Srivastava G, Srivastava R. A survey on automatic image captioning [C]// Proc of International Conference on Mathematics and Computing. Berlin, Germany: Springer Press, 2018: 74-83.

[117] Wu Qi, Teney D, Wang Peng, *et al.* Visual question answering: a survey of methods and datasets [J]. Computer Vision & Image Understanding, 2017, 163: 21-40.

[118] Seong T W, Ibrahim M Z. A review of audio-visual speech recognition [J]. Journal of Telecommunication, Electronic and Computer Engineering, 2018, 10 (1-4): 35-40.

[119] Gehring J, Auli M, Grangier D, *et al.* Convolutional sequence to sequence learning [C]// Proc of the 34th International Conference on

Machine Learning. New York: ACM Press, 2017: 1243-1252.
[120] Yuxin Peng, Xin Huang, Jinwei Qi. Cross-media shared representation by hierarchical learning with multiple deep networks [C]// Proc of the 25th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Press, 2016: 3846-3853.